University of Alberta

COMPARISON OF ABILITY ESTIMATES USING UNIDIMENSIONAL AND MULTIDIMENSIONAL SCORING MODELS FOR DICHOTOMOUSLY SCORED ITEMS

by

NIZAM RADWAN

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy

in

Measurement, Evaluation, and Cognition

Department of Educational Psychology

Edmonton, Alberta

Fall 2005

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-08719-6

Your file *Votre référence*
ISBN:
Our file *Notre reference*
ISBN:

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Dedication

To

Sanaa, Malak, and Farah

for their love.

Abstract

Unidimensional scoring models have been used to score multidimensional tests because of the lack of availability of computer programs to conduct multidimensional item response model analyses. The purpose of this study was to assess the degree to which the classical test score model (CTM) and the 2-parameter unidimensional item response model with single calibration of all items (UIRM (T)) and separate calibration of the items within the subtest (UIRM (S)) were able to recover (a) the scores obtained by and (b) the classifications made using the multidimensional 2-parameter compensatory item response model (MIRM) when the dimensionality of the test was known to be two.

Simulated data were generated where the same samples of examinees were used across the scoring methods for each condition. The factors considered were the correlation between examinees abilities on the two dimensions, the mean differences on the two dimensions, the factor complexity, and the type of score reported. The agreement between the scores yielded by the MIRM scoring model and each of the remaining three scoring models was assessed by (a) the differences between correlations between examinees' scores, (b) the differences between the correlations between the examinees' subtest scores within scoring method, (c) the root mean square difference (RMSD) between examinees' scores, and (d) the differences in rates of correct and incorrect classification of examinees.

All scoring methods ranked examinees similarly. The recovery of the correlations between the examinees' subtest scores was complex across and within scoring method. The RMSD was small except when the subtest scores were used and the test structure

was complex. The classification results revealed high rates of agreement based on the mean percentage between the MIRM and each of the CTM, UIRM (T), and UIRM (S). Taken together, the results suggest that the use of multidimensional, unidimensional, and number-right scoring will not lead to differences when total test scores derived from multidimensional tests with simple structure are reported. The results are equivocal with subtest scoring when the structure is complex.

Acknowledgements

I would like to express my appreciation to many people whose help and guidance have made the completion of this dissertation possible.

I am especially grateful to Dr. W. Todd Rogers for accepting to be my academic supervisor and mentor. I am forever in debt to Dr. Rogers for his patience, support, and advice on academic and non-academic issues throughout my doctoral program. A special thanks to Dr. Mark J. Gierl for introducing me to item response theory and for his support and encouragement to pursue this thesis.

A great appreciation goes to Dr. Michael D. Carbonaro, Dr. Stephen P. Norris, Dr. Cameron Wild, and Dr. Cindy Walker for their willingness to serve on my thesis committee. Their support and constructive feedback greatly improved the quality of this work.

I would also like to thank Dr. Colin Fraser for his help with the program NOHARM III, and Dr. R.P. McDonald for his help on issues related to factor analysis. A special thanks to my colleagues at CRAME for their support, especially Xuan Tan for her assistance with syntax.

Table of Contents

CHAPTER 1: INTRODUCTION 1

BACKGROUND INFORMATION 1

PURPOSE OF RESEARCH 4

Rationale 5

DEFINITION OF TERMS 6

ORGANIZATION OF THE THESIS 7

CHAPTER 2: LITERATURE REVIEW 8

OVERVIEW 8

DIMENSIONALITY 9

CLASSICAL TEST SCORE MODEL 13

Model 14

Assumptions 14

Score Calculation 17

Advantages and Limitations of Scoring using Classical Test Model 18

UNIDIMENSIONAL ITEM RESPONSE MODELS 19

Models 20

Assumptions 22

Score Estimation 23

Advantages and Limitations of Scoring using UIRMs 25

MULTIDIMENSIONAL ITEM RESPONSE MODELS 26

Models 26

Assumptions 31

Score Estimation 32

Advantages and Limitations of Scoring Using MIRMs 33

REVIEW OF PREVIOUS RESEARCH STUDIES 34

Previous Research: UIRM 34

Previous Research: MIRM 39

Previous Research: Classification 43

Summary of the Limitations of Previous Research 45

SUMMARY 45

CHAPTER 3: METHOD **47**

OVERVIEW 47

METHOD 47

Data 47

Multidimensional Item Parameters 48

Simple Versus Complex Structure 52

Mean of Ability Distributions 53

Correlation Between Abilities 55

Design 55

PROCEDURE 56

Data Generation 56

Classical Test Model Scoring Model 57

Ability Estimation 57

Unidimensional Item Response Scoring Models 58

Multidimensional Item Response Scoring Model 58

Total Test Score Versus Subtest Scores 60

ANALYSES 60

Correlation 61

Correlation Between Examinees' Subtest Score 62

Root Mean Square Difference 62

Classification 63

CHAPTER 4: RESULTS AND DISCUSSION 65

PRELIMINARY ANALYSES 65

CORRELATIONAL AGREEMENT 68

CORRELATIONS BETWEEN SUBTESTS WITHIN SCORING METHOD 70

ROOT MEAN SQUARE DIFFERENCE 72

Total Score 72

Subtest scores 75

CHAPTER 5: SUMMARY AND CONCLUSIONS 85

SUMMARY OF THE STUDY 85

SUMMARY OF KEY FINDINGS 86

LIMITATIONS OF THE STUDY 89

CONCLUSIONS 89

REFERENCES 94

APPENDIX A 100

TOTAL SCORE 100

APPENDIX B 122

SUBTEST SCORES 122

APPENDIX C 151

MULTIDIMENSIONAL ITEM STATISTICS 151

List of Tables

Table 1: *LSAT Item Parameters for October 1992 Administration.* 49

Table 2: *Item Parameters for Simple and Complex Structure.* 51

Table 3: *Proportion of Explained Variance for Linear and Cubic Functions.* 68

Table 4: *Mean Correlations for Total Test Score.* 69

Table 5: *Mean Correlations for Subtest Scores.* 70

Table 6: *Mean Correlations between Subtests within Scoring Methods.* 71

Table 8: *Root Mean Square Difference for Subtest Scores.* 76

Table 9: *Classification of MIRM and CTM, UIRM (T), and UIRM (S) for Total Score - Simple Structure.* 81

Table 10: *Classification of MIRM and CTM, UIRM (T), and UIRM (S) for Total Score – Complex Structure.* 82

Table 11: *Classification of MIRM, CTM, and UIRM (S) for Subtest Scores – Simple Structure.* 83

Table 12: *Classification of MIRM, CTM, and UIRM (S) for Subtest Scores – Complex Structure.* 84

List of Figures

Figure 1. Simple Structure 12

Figure 2. Complex Structure 13

Figure 3. Item Characteristic Curve: 2-Parameter Model. 21

Figure 4. Item Characteristics Surface. 28

Figure 5. Contour Plot of Item Characteristic Surface. 29

Figure 6. Vector Plot of Relatively Low and High Discriminating Items 30

Figure 7. Ability Distributions with Mean Vectors (0, 0). 54

Figure 8. Ability Distributions with Mean Vectors (0, -1). 55

Figure 9. Cut – Score for Dimensions 1 and 2. 64

Figure 10. Scatter Plot of CTM and MIRM – 0000ST 67

Figure 11. RMSD for Total Score – Simple Structure. 74

Figure 12. RMSD for Total Score – Complex Structure. 75

Figure 13. RMSD for Sub-Test Scores – Simple Structure 77

Figure 14. RMSD for Sub-Test Scores – Complex Structure. 78

CHAPTER 1: INTRODUCTION

Background Information

Educational tests have increasingly gained importance in our society. Educational policies and examinees' educational progress are dependent on tests and their results. Two kinds of educational tests are used to measure the knowledge of examinees and to form educational policies: classroom testing and large-scale testing. As the name suggests, classroom testing is administered in schools where teachers develop the subject area tests to measure each of their student's standing relative to other students in their classes, or to measure the amount of knowledge that the students mastered in each subject area. Provincial and the federal governments, however, administer large-scale tests to measure examinees' achievement, hold schools accountable, and adjust educational policies. Norm-referenced or criterion-referenced test score interpretations, or, in some jurisdictions, both norm- and criterion-referenced test score interpretations are made. The scoring of large-scale tests was addressed in this study.

The ultimate goal of testing is to obtain a measure of the knowledge and skills examinees possess. The measure is quantified as a test score that is used to provide information for making decisions, such as selection, certification, placement, and diagnostic decisions. The scores yielded by the test must be such that it is possible to draw inferences where "the inferences drawn about the knowledge, skills, attitudes, and behaviors possessed by each student are valid and not open to misinterpretation" (*Principles for Fair Student Assessment Practices for Education in Canada*, 1993, p. 3). The accuracy of these scores is very important to create a fair assessment that leads to

valid inferences about the examinees' knowledge, skills, or ability. According to the *Standards of Educational and Psychological Testing*, validity refers to "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (*American Educational Research Association, American Psychological Association, & National Council on Measurement in Education*, 1999, p. 9).

Test scores are determined from items that typically are scored right or wrong. While there has been a resurgence of the use of polytomously-scored items, the majority of items are dichotomously scored. Hence, the focus of this proposed study was on dichotomous items and how they should be scored. Several models for scoring dichotomous data are available. These models include the classical test score model (Spearman, 1904), unidimensional item response models (Lord, 1952; Rasch, 1960), and multidimensional compensatory item response models (McKinley & Reckase, 1982). Despite the differences among these models in terms of how the scores are calculated, they all share the common goal of optimally providing evidence based on examinees responses to items to infer an examinee's level of performance (Thissen & Wainer, 2001).

The classical test score model and the unidimensional item response models are based on the underlying assumption that all the items in a test measure only one "dimension." However, the nature of the construct measured may or may not be unidimensional. For example, if all the items measure the same combination of two different skills (e.g., verbal ability and computational ability) then the test is considered unidimensional despite the fact that the construct is multidimensional. In this case, obtaining a total test score using the classical test score model or one of the

unidimensional item response models may be appropriate even though the dimensions are distinct. Scoring methods have recently been developed using multidimensional item response models to score multidimensional data. However, due to the complexity of the procedures involved in obtaining a total score or subtest scores using unidimensional or multidimensional item response scoring models, the number-right scoring method remains the most popular and widely used procedure (Ndalichako & Rogers, 1997).

Many testing programs establish performance standards and the corresponding cut-scores in the distribution of scores. For example, students can be classified as masters or non-masters. The unidimensional and the multidimensional scoring models may lead to classifying examinees differently when the data is known to be multidimensional. For example, one scoring model may classify an examinee as a non-master whereas another scoring model may classify the same examinee as a master. Consequently, this study examined the effect of using different scoring models to assess their differences in classifying examinees to different categories when the data are known to be two-dimensional.

Several researchers, such as Anderson (1999), Fan (1998), Ndalichako and Rogers (1997), Rogers and Ndalichako (2000), and Tomkowicz and Rogers (in press), compared the classical test score model and the unidimensional item response scoring models for dichotomously-scored items and found that the scores yielded by these models were similar. In the case of multidimensional data, Luecht (2003) compared the classical test score model, the 3-parameter unidimensional item response model, and the 3-parameter compensatory multidimensional item response model and found that the unidimensional item response model in which the subtest items were calibrated

separately yielded similar scores to the multidimensional model. Tate (2004) investigated the effects of *maximum likelihood* and *expected a posteriori* estimation methods used in item response models on the total score and subtest scores. He found that the estimation procedures performed differently under different conditions. Walker and Beretvas (2003) examined the effects of using the unidimensional 3-parameter model and the multidimensional compensatory 3-parameter model on the classification of examinees into four proficiency levels. They found that examinees who had low ability in mathematical communication were more likely to be classified as less proficient on general mathematical ability when the unidimensional scoring model was used compared to the multidimensional scoring model. However, a comprehensive comparison of the scores yielded by different scoring models under different conditions when the construct is known to be two-dimensional and their effect on the classification of examinees is lacking.

Purpose of Research

The purpose of the present study was to assess the degree to which the classical test score model (CTM), the unidimensional 2-parameter item response model with single calibration of all items (UIRM (T)) and separate calibration of items for each subtest (UIRM (S)) were able to recover (a) the scores obtained by and (b) the classification made using the multidimensional 2-parameter compensatory item response model (MIRM) when the dimensionality of the test was known to be two. Since the 2-parameter, 2-dimensional compensatory item response model is the appropriate model for the multidimensional data, the scores yielded by this model were considered the "true" scores.

To address the purpose of this study, a simulation study was conducted. The factors considered included:

1. the correlation between examinees abilities on the two dimensions: 0.0, 0.3, 0.6, and 0.9;
2. mean differences on the two dimensions: equal (0,0) and unequal (0, -1);
3. factor complexity: simple and complex; and
4. type of score reported: total score and subtest score.

Rationale

This proposed study was based on the following rationale. Since test scores are used as evidence to make important decisions about examinees, the scores must be determined accurately in order for sound decisions to be made. Inaccurate total test scores and subtest scores can adversely influence the inferences made about the performance of examinees. A comparative evaluation of the accuracy of the classical test score model and the unidimensional test score model when the data are known to be two-dimensional will allow decision-makers to choose the scoring method that is most accurate for the test of interest. This choice will lead to valid inferences about examinees' knowledge or skills and may lead to classifications of examinees into different categories that are more accurate.

This research study was conducted as part of a program of research at the Center of Research in Applied Measurement and Evaluation (CRAME) and it extends the previous work of Ndalichako and Rogers (1997), Rogers and Ndalichako (2000), and Tomkowicz and Rogers (in press) to the multidimensional case.

Definition of Terms

Dichotomous scoring: scoring procedure where "1" is given for a correct response and "0" is given for an incorrect response to test items.

Dimensionality: a term that broadly refers to the number of dimensions the test is measuring.

Classical test score model: a model that specifies the relationship between an observed score, true score, and error of measurement.

Unidimensional item response models: models that specify the relationship between an underlying ability (θ) and the probability of an examinee answering the item correctly. These models differ in terms of the number of item parameters to be estimated: difficulty, discrimination, and/or pseudo-chance.

Compensatory multidimensional item response models: models that specify the relationship between two or more underlying abilities and the probability of an examinee answering the item correctly. The compensatory nature of this model allows an examinee that has a low ability on one dimension to compensate for it by having a high ability on a second dimension.

Simple structure: the items related to a dimension in a multidimensional space in a test measure primarily that dimension.

Complex structure: the items measure two or more of the dimensions in a multidimensional space of a test to a certain extent.

Classification: Assigning examinees to groups based on established standards and their corresponding cut-scores.

Organization of the Thesis

Issues relating to scoring dichotomously scored items and a presentation of the research purpose and its rationale have been discussed followed by the definition of terms in Chapter 1. A review of dimensionality, the classical test score model, the unidimensional item response models, and the multidimensional item response models is presented in Chapter 2. The chapter concludes with a critical review of research studies in which the different scoring methods for dichotomously scored items were compared when the data were unidimensional and when the data were multidimensional. A detailed account of the methods used to answer the research questions along with the computer programs and scoring procedures used is presented in Chapter 3. This is followed by the presentation and discussion of the results in Chapter 4. The summary of the study and its results, the limitations, the conclusions drawn, the implications for future practice, and the implications for future research are presented in Chapter 5.

CHAPTER 2: LITERATURE REVIEW

Overview

Test scores play an important role in assessing the ability, skills, or knowledge of examinees because they provide information that is used for making decisions that affect examinees' futures, such as selection to pursue higher levels of education, job promotion, and certification and licensure. Thissen and Wainer (2001) defined the test score as a "summary of the evidence contained in an examinee's responses to the items of a test that are related to the construct or constructs being measured" (p. 1). Test scores are usually reported as total scores or as sub-scores for each of the content areas that are assessed by the test. Regardless of the number of content areas assessed, the total test score is the most prevalently used score and, because of this, is most familiar to educators and the general public. However, reporting a total score may not be appropriate under certain conditions, such as when the test is multidimensional and the dimensions are weakly correlated. In this case, the use of subtest scores may be more correct than the use of the total test score. Since test scores are used as the basis for making important decisions about the examinees and since most of the educational tests are hypothesized to be multidimensional (Ackerman, 1994), the evaluation of the accuracy of these scores when the data are multidimensional becomes paramount. The accuracy of the total score and the sub-scores can affect examinees, especially when the decision is pass/fail or accept/reject.

The literature review is organized in three sections. The first section includes an overview of dimensionality. The second section includes a review of the models that form the basis of the scoring methods used in this study: the classical test score model,

the unidimensional item response models, and the multidimensional item response models. In the third section, a review of the studies in which different scoring methods were compared when the data were unidimensional and in other cases when the data were multidimensional is presented.

Dimensionality

The dimensionality of a test refers to the number of dimensions that the test is measuring. Reckase (1990) distinguished between psychological and statistical dimensionality. Psychological dimensionality refers to the number of hypothesized psychological constructs needed for an examinee to successfully perform on a test (Embretson, 1985; Reckase, 1990). For example, "numerical computation and verbal reasoning are said to be required to successfully perform on a mathematics story problem" (Reckase, 1990, p. 2). In contrast, statistical dimensionality refers to the minimum number of dimensions required to summarize the examinees' data matrix. For example, "a vector composed of two elements may be needed in a probabilistic model of test performance to reasonably accurately predict how a person will respond to a particular set of test items" (Reckase, 1990, p.2).

Formally defined, the statistical dimensionality of a test is the minimum number of dimensions required to produce a latent model that is both locally independent and monotone (Stout, 1990; Tate, 2002). Local independence is defined as, "the condition that the probability of any pattern of responses to all of the items, conditioned on the abilities, is equal to the product of the conditional probabilities of each of the responses" (Tate, 2002, p. 184). A monotone model means a model in which "the probability of a

correct item response monotonically increases with increasing values of the dimensions" (Tate, 2002, p. 184).

However, the above definition of statistical dimensionality is strict because it implies that a unidimensional test has only one dimension that accounts for examinee performance. In reality, however, more than one dimension may affect, to a certain extent, the examinees performance on a test (e.g., test anxiety, guessing, and speed) (Hambleton *et al.* 1991). Similarly, the concept of local independence requires that the covariance between a pair of test items equals zero at each ability level. In practice, however, this assumption cannot be strictly met because even after conditioning on ability, a small covariance between a pair of items may still be present.

Stout (1990) proposed that only the number of major or dominant latent dimensions in a test should be considered and any minor dimensions could be ignored. This approach to determining the dimensionality of a test is termed "essential dimensionality." Essential dimensionality is based on essential local independence -- a weaker version of local independence. Essential local independence requires that the covariance between all pairs of item responses, conditioned on ability, be small in magnitude (Stout, 1990). Based on this definition of dimensionality, a test is demonstrated to be essentially unidimensional if one dominant dimension accounts for examinee performance on the test. In contrast, a test is demonstrated to be multidimensional if more than one dominant dimension accounts for examinee performance on the test.

It should be noted, however, that the dimensionality of the test is not only due to test items, but also to the interaction between examinees and test items (Reckase, 1990;

Ackerman, 1994; Tate, 2002). A test structure may be multidimensional but the resulting data need not be multidimensional. A test that has a multidimensional structure could result in unidimensional data if the performance of the population of examinees is homogeneous with respect to all of the dimensions assessed by a test. For example, in a multidimensional test made up of multiple algebra story items, the response data would be unidimensional if the target population of examinees was homogenous with respect to level of algebra knowledge and skill and reading required by all the items (Tate, 2002). If all the items in a test measure the same composite of abilities in the same way, then the resulting response data will also be unidimensional (Reckase, Ackerman, & Carlson, 1988; Reckase, 1990). Unidimensional data will also result if the psychological dimensions of a test are strongly confounded with the difficulty of the test items. This case will result in a unidimensional data structure because "there is little variation in the probability of correct response on items measuring other dimensions when there is little variation in the probability of correct response for items measuring the first dimension" (Reckase, 1990, p. 25).

If a test structure is demonstrated to be unidimensional, then reporting the total score is appropriate because all of the items in the test are measuring one dominant ability. The total test score represents all test-related differences among examinees. When the test structure is demonstrated to be multidimensional, the total score can also be viewed as a composite of different abilities. If the subtest scores are used in addition to the total score, then the subtest scores are assumed to represent lower levels of abilities or specific abilities related to the different components of the test whereas the total score is assumed to represent a higher level of ability or a general ability. For example, in a

mathematics test that is comprised of algebra, geometry, number concept, and probability items, the subtest scores represent a specific ability for each of the test components whereas the total score represents general mathematics ability (Tate, 2002). In this case, the correlation among the mathematics subtests will be high in comparison to a situation in which the correlation among the subtests is low or weak. In this latter case, the use of a total score may not be warranted.

When a test is multidimensional, its structure in a multidimensional space can be simple or complex (Stout *et al.* 1996). In a Cartesian coordinate system, a test exhibits simple structure if the vectors of all of the items referenced to a particular dimension lie along or close to the corresponding coordinate axis. The two-dimensional test shown in Figure 1 exhibits simple structure because the four item vectors lie along or close to the two coordinate axes. The vectors for items 1 and 2, which assess the first dimension lie close to axis D1, thus measuring dimension 1well. The vectors for items 3 and 4, which assess the second dimension lie close to axis D2, thus measuring dimension 2 well.

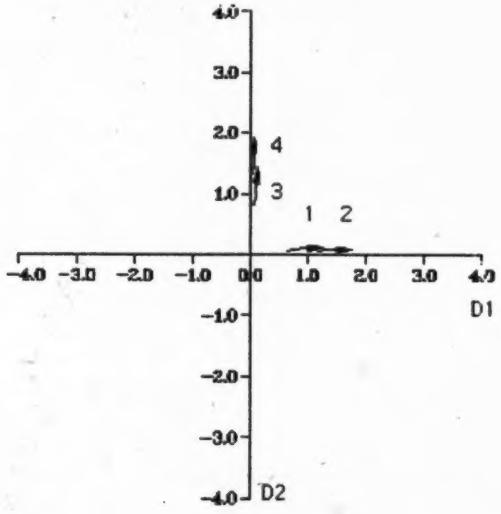


Figure 1. Simple Structure

A test exhibits complex structure if the item vectors lie between the coordinate axes. The two-dimensional test shown in Figure 2 exhibits complex structure because the four item vectors lie between the two coordinate axes, thus measuring dimensions 1 and 2 to various degrees.

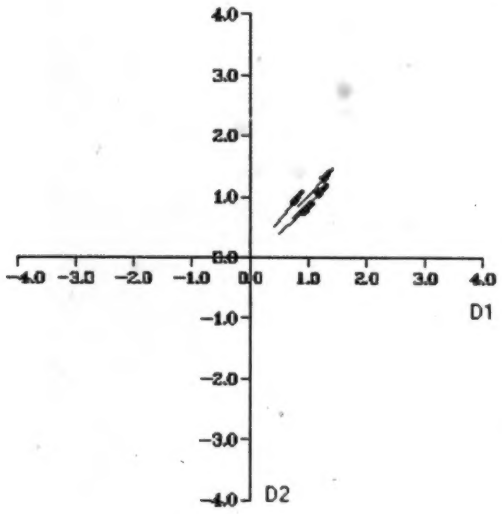


Figure 2. Complex Structure

Classical Test Score Model

Spearman (1904) proposed the earliest model of measurement. Known as the classical test model (CTM), this model is based on the assumption that each examinee has a true score. However, the measurement process always contains errors to some degree. The major task is to estimate the relationship between the observed score and the true score.

Model

The model states that an examinee's observed score, X_{jf}, is composed of two additive components:

$$X_{jf} = \tau_j + \varepsilon_{jf},$$

where

X_{jf} is the observed score of examinee j on form f,

τ_j is examinee j 's true score, and

ε_{jf} is the error score of examinee j on form f (Lord & Novick, 1968, p. 56).

The difference between the examinee's observed score and his/her true score is defined as the error of measurement. It should be noted that the true score and the error score are not observable. Therefore, to estimate the relationship between the true score and the observed score, a set of assumptions must be made.

Assumptions

Eight assumptions are made in the classical test score model. The first three assumptions relate to an examinee and the last five assumptions relate to the population of examinees.

A. Individual Examinee

1. The previous equation can be regarded as an assumption that states the examinee's observed score, true score, and error score are linearly related:

$$X_{jf} = \tau_j + \varepsilon_{jf}.$$

2. The expected observed score is defined as the examinee's true score:

$$\xi(X_{jf}) = \tau_j, f \to \infty,$$

where $\xi(X_{jf})$ is the expectation of examinee j 's observed score across parallel forms f.

Lord and Novick (1968, p. 154) stated that "by definition, an unbiased estimator of a parameter is an estimator whose expected value is the parameter estimated." Therefore, assumption 2 indicates that the observed score is an unbiased estimate of the true score. Consequently, the expected value of an examinee's error of measurement is zero.

3. The error of measurement for an examinee j across an infinite number of parallel forms f is assumed to be a random and normally distributed variable:

$$\varepsilon_{jf} \approx NID(0,\sigma_{\varepsilon}^2),$$

where σ_{ε}^2 is the variance of the error of measurement for examinee j across trials f.

B. Population of Examinees

4. The mean of errors of measurement over the population of examinees on form f is zero:

$$\xi_j(\varepsilon_{jf}) = 0.$$

5 and 6. The errors of measurement for examinees on form f are random and normally distributed:

$$\varepsilon_{jf} \approx NID(0,\sigma_e^2),$$

where σ_e^2 is the variance of the error of measurement for the population of examinees across trials f.

7. The correlation between true and error scores is zero in the population of examinees:

$$\rho_{\tau e} = 0.$$

8. The correlation between the errors on two parallel forms in the population of examinees is zero:

$$\rho_{e_1 e_2} = 0.$$

To be parallel, it is assumed that the two forms of the test are interchangeable. In other words, the means and the variances of the observed scores obtained from the two forms are equal and the correlation between their observed scores is equal to one. In this case, the two forms are strictly parallel. Since satisfying the condition of parallelism is difficult, a less restrictive definition of parallel forms requires that only the means of the observed scores obtained from the two forms are equal (Rulon, 1939; Guttman, 1945; and Flanagan, (Kelly, 1942)). In addition, the forms need to be relevant and representative of the same construct.

Unfortunately, it is not possible to obtain unique estimates of the true score and the error score at the individual examinee level because it is not feasible to assess an examinee an infinite number of times with the same test or with an infinite number of parallel forms of the test (Rogers, 2000). Consequently, the variance error of measurement in the population of examinees is used to estimate the variance error of measurement for the individual examinee. Spearman (1904) showed that in a population of examinees:

$$\sigma_X^2 = \sigma_\tau^2 + \sigma_e^2,$$

where

σ_X^2 is the observed score variance across examinees,

σ_τ^2 is the true score variance across examinees, and

σ_e^2 is the variance of the error of measurement across examinees.

The variance of the error of measurement across examinees is the mean of the individual examinee variance in the population:

$$\sigma_e^2 = \frac{\sum_{j=1}^{N} \sigma_{\varepsilon_j}^2}{N}, N \to \infty,$$

where

$\sigma_{\varepsilon_j}^2$ is the variance of the error of measurement for the j's examinee, and

N is the number of examinees (Lord and Novick, 1968, p. 155).

When the variance of the error of measurement across examinees is close to zero, we can infer, based on the above formula, that the variance of the error of measurement for the individual examinees is small (Rogers, 2000).

Score Calculation

Total Score

The total score is obtained by summing the examinees' correct responses across the test items. The items in this case are assumed to have equal weights; i.e., the weight for each item equals to 1. However, some items may be assigned more weight than other items. In this case, the observed score equals the weighted sum of correct responses across test items:

$$X_j = \sum_{i=1}^{I} w_i X_{ij},$$

where

X_j is the observed score for examinee j,

w_i is the weight of item i,

X_{ij} is the examinee j response to item i, and

I is the number of items.

Sub-scores

The total score could also be viewed as the sum of the scores of each content area. For example, the Law School Admission Test (LSAT) has three content areas: analytical reasoning, logical reasoning, and reading comprehension. If scores are obtained for each of the content areas, then these scores are called sub-scores. The total test score in this case is the sum of the sub-scores of the content areas. In other words, the total score is a composite score:

$$X_{c_j} = \sum_{v=1}^{m} w_v X_{vj},$$

where

X_{cj} is the composite score for examinee j,

w_v is the weight for component v, and

X_{vj} is the score of examinee j on component v.

When the test is multidimensional, the sub-scores are used in addition to the total score to provide more information about examinees performance.

Advantages and Limitations of Scoring using Classical Test Model

The main advantage of computing the total observed score and, if appropriate, subtest scores using the CTM is the simplicity of the CTM. Adding up the number of items that the examinee answered correctly or computing the weighted total score are relatively simple. Another advantage of scoring based on classical test score model is the

ease of explaining the process to the general public. Educators and the general public are much more familiar with this scoring method than with scoring based, for example, on an item response model (Rogers & Ndalichako, 2000).

However, one of the limitations of the CTM is that the examinees' scores are test dependent. The examinees' scores depend on the difficulty level of the test items. For example, examinees will score low on a difficult test and will score high on easier test with the same content (Hambleton *et al.* 1991). Therefore, the examinees scores will be high or low depending on the difficulty of the test items.

Another limitation of the classical test score model is that the standard error of measurement is usually assumed to be the same for all examinees. This assumption is problematic because the standard error of measurement tends to differ at different ability levels. For example, examinees who have low scores tend to have larger errors of measurement relative to other examinees. The larger error of measurement for the examinees with low scores is partially due to guessing.

Unidimensional Item Response Models

The unidimensional item response models (UIRMs) provide a different approach to score estimation than the classical test score model. Item response models take into account the interaction between examinees and test items. It is assumed that there is an underlying ability (θ) that influences this interaction. Examinees who have high levels of ability have higher probabilities of answering an item correctly than examinees with lower levels of ability. The monotonically increasing function used to represent the relationship between ability and the probability of responding correctly to a test item

across the ability scale is known as the item characteristic curve (ICC) (Hambleton et al., 1991).

Models

Three UIRMs have been most widely used with dichotomously-scored items: the one-, two-, and three-parameter models. As their names suggest, the primary distinction between the three models is the number of item parameters needed to account for the performance of examinees.

Since the two-parameter logistic model was the model used in this study, it is presented first. The two-parameter logistic model is given by:

$$P(X_{ij} = 1|\theta_j) = \frac{e^{1.7a_i(\theta - b_i)}}{1 + e^{1.7a_i(\theta - b_i)}}, \quad i = 1,2...I,$$

where

$P(X_{ij} = 1|\theta_j)$ is the probability that a randomly chosen examinee j with ability θ answers item i correctly,

a_i is the item discrimination parameter for item i,

b_i is the difficulty parameter for item i (the point on the theta scale where the probability of answering the item correctly is 50%),

I is the number of items in the test, and

e is a transcendental number whose value is 2.718 (corrected to three decimal places).

It should be noted that the above equation represents the logistic model as opposed to the normal ogive model. The logistic model is commonly used in item

response models because it is more mathematically tractable than the normal ogive model.

The parameter estimated in the one-parameter logistic model is the *b* parameter. The items are assumed to be equally discriminating and the examinees do not guess. For the three-parameter logistic model, the parameters estimated are the *a*-, *b*-, and *c*-parameters. The examinees are assumed to guess when they do not know the correct answer. However, the value of the *c*-parameter is often not equal to the value of an examinee randomly guessing; therefore, the *c*-parameter is called the *pseudo-chance* parameter. For the three-parameter logistic model, the *b* parameter is the point on the ability scale where the probability of correct response is equals to $\frac{c_i + 1}{2}$. Figure 3 shows a 2-parameter item characteristic curve with a difficulty parameter of 0.11 and a discrimination parameter of 0.71.

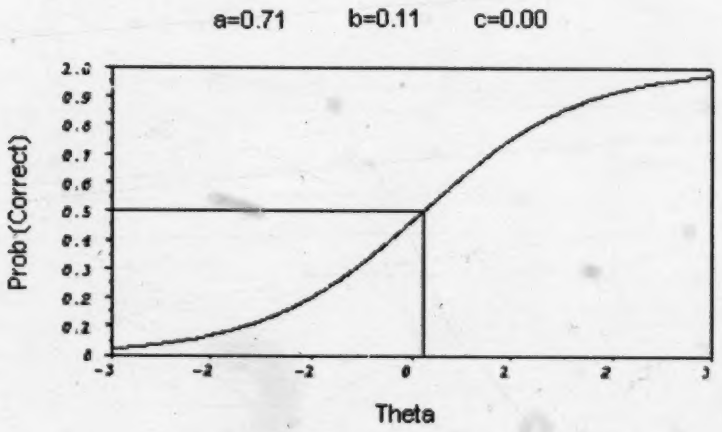


Figure 3. Item Characteristic Curve: 2-Parameter Model.

Although more restrictive than the 3-parameter model, the 2-parameter model was used in this study to estimate items and ability parameters. The reason is that the pseudo-chance or *c*-parameter in the 3-parameter model may have detrimental effect on the estimation of ability. Generally speaking, the effect of the *c*-parameter is to lower the amount of information an item provides. All else being equal, the 1PL or 2PL models are more informative than the 3PL model (Embretson & Reise, 2000, p.184).

Assumptions

Four assumptions are made about the data to which the item response model is applied: a true relationship between ability and item response exists, unidimensionality, local independence, and non-speededness of response.

1. The item characteristic curve reflects the true relationship between the underlying ability and the examinees item responses.
2. The assumption of unidimensionality states that all items in a test measure one underlying ability. In other words, one ability is needed to account for the examinee performance on a test. The assumption cannot be strictly met because of several factors that affect examinee performance (Hambleton *et al.*, 1991). The assumptions of unidimensionality and essential unidimensionality have been previously discussed in the dimensionality section of this chapter.
3. Local independence means that after controlling for ability, the responses of examinees to any pair of items are independent. In other words, the ability of the examinee is the only factor influencing their responses to the test items (Hambleton, et al. 1991). This definition of local independence is known as the strong local independence assumption. A weaker version of local independence,

essential local independence, has been previously discussed in the dimensionality section. Local independence is related to the assumption of unidimensionality. When the assumption of unidimensionality is met, the assumption of local independence is also met (Lord & Novick, 1968).

4. The non-speededness of response assumption means that the examinees have sufficient time to respond to all the test items. Therefore, the test is considered a power test and not a speeded test. In a power test, the examinees must attempt at least 85% of the items.

Score Estimation

Three score estimation procedures are mainly used in the UIRMs for dichotomously-scored items: *maximum likelihood estimation* (MLE), *modal a posteriori* (MAP), and *expected a posteriori* (EAP). The MLE is a method for estimating examinee ability where the likelihood of an examinee response with a particular pattern is the product of the probabilities of observing the response to each item. The ability at which the likelihood function of an examinee is a maximum is considered the best ability estimate of that examinee. The MAP is a Bayesian estimation procedure where a *prior* probability distribution is used in addition to the likelihood function to obtain the *posterior* probability distribution in which the prior probabilities are multiplied by the likelihood function. The mode of the resulting posterior distribution is considered an examinee's best ability estimate. Similar to the MAP, the EAP is a Bayesian ability estimation procedure where the mean, instead of the mode, of the posterior distribution is considered the best estimate of an examinee's ability. Since the score estimation

procedure used in this study was *expected a posteriori*, the review was restricted to this procedure.

Compared to the other estimation procedures, the EAP has the following advantages over MLE and MAP (Bock & Mislevy, 1982). First, the EAP is a non-iterative process and thus is computationally faster. Second, unlike MAP, EAP does not require the calculations of derivatives of the response function because it employs a discrete prior distribution. Third, unlike MLE, EAP estimates exist for examinees that respond to all items correctly, examinees that respond to all items incorrectly, and examinees with aberrant response patterns. Fourth, the EAP estimates have the smallest mean square error over the population specified by the prior distribution. However, one disadvantage of using EAP is that the ability estimates regress to the mean ability unless the number of items is large (Wainer & Thissen, 1987; Embretson & Reise, 2000).

Expected A Posteriori

As indicated in the preceding section, the *expected a posteriori* is a Bayesian estimation method where the mean of the posterior distribution is calculated to estimate the examinee ability. To obtain the mean, a number of quadrature points on the ability scale are specified. The quadrature points are equally spaced across the range of the ability scale. The weights are set equal to the prior discrete probability at these quadrature points. The prior probability distribution can be based on a probabilistic model, belief, or experience (Suen, 1990). The prior probability distribution is assumed to be normal in most cases. The posterior distribution is obtained by accumulating the posterior probabilities over the subjects at each quadrature point. The sums are then normalized so an estimate of the probabilities at the points can be obtained. The

posterior probability distribution is an updated distribution of ability after the data from a sample have been considered. Mislevy and Bock (1990) suggest that $2\sqrt{I}$ (where I is the number of items) as a rule of thumb to specify the number of quadrature points for a large number of items. The *expected a posteriori* (Bock & Mislevy, 1982) is given by:

$$\overline{\theta_j} = \frac{\sum_{k=1}^{q} X_k L_j(X_k) W(X_k)}{\sum_{k=1}^{q} L_j(X_k) W(X_k)}$$

where

X_k	k is the k^{th} quadrature point, $k=1...q$,
$L_j(X_k)$	is the likelihood of examinee j at quadrature point k,
$W(X_k)$	is the weight associated with the k^{th} quadrature point, and
q	is the number of quadrature points.

The weights are the probabilities at the corresponding points of the discrete prior distribution. The weights are normed so that

$$\sum_{k=1}^{q} W(X_k) = 1.$$

Advantages and Limitations of Scoring using UIRMs

One of the advantages of scoring using UIRMs is that the ability parameter estimate is invariant across test forms. Unlike the classical test model score estimation, the invariance of the ability estimates means that the examinees' scores or ability are not test-dependent. This property is the cornerstone of UIRMs and is their major advantage over the classical test score model (Hambleton et al., 1991).

Another advantage is that the error of measurement is not the same across the score scale. The errors of measurement tend to be higher for examinees with low and

high ability than for examinees with average ability. Therefore, the UIRMs allow for different precision measurements of ability estimates at different points on the score scale.

However, score estimation using the UIRMs requires large sample size in order to obtain stable ability and item parameter estimates. The need for large sample size restricts the use of UIRMs to large-scale assessments. In addition, the complexity of score estimation requires the use of high-speed computers and statistical knowledge. This is a downside of estimating scores using the UIRMs because only individuals with specialized training are able to obtain and interpret the estimated scores. Finally, since one of the assumptions is the presence of one dominant ability to account for the examinee performance, the UIRMs may not model the data accurately when the data are multidimensional.

Since most of educational tests are multidimensional, extensions of the UIRMs have been developed to address the issue of multidimensionality. The score estimation procedures used in UIRMs have been extended to the multidimensional case.

Multidimensional Item Response Models

Multidimensional item response models (MIRMs) are used to model the interaction between examinees and items when the assumption of unidimensionality is not met. MIRMs model the interaction between examinees and test items when there are two or more dimensions.

Models

Two types of MIRMs can be used to describe the data of dichotomously-scored items: the compensatory model and the non-compensatory model. Estimation methods

are currently available for the multidimensional compensatory model; therefore, the review is restricted to this model. Since the multidimensional two-parameter compensatory model was the model used in this study, it will be presented first. The two-parameter logistic compensatory model is given by (Ackerman, 1996):

$$P(X_{ij} = 1 | \mathbf{a}_i, d_i, \boldsymbol{\theta}_j) = \frac{e^{(1.7\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}}{1 + e^{(1.7\mathbf{a}_i'\boldsymbol{\theta}_j + d_i)}},$$

where

X_{ij} is the score on item i by person j,

$\mathbf{a}_i'$ is the transpose of the $m \times 1$ vector of item discrimination parameters; m is the number of item discrimination parameters corresponding to the number of dimensions,

d_i is a scalar parameter related to the difficulty of item i,

$\boldsymbol{\theta}_j$ is the $m \times 1$ vector of ability parameters for person j.

The item parameter estimated in the multidimensional one-parameter logistic compensatory model is the *d*-parameter. The items are assumed to be equally discriminating across the dimensions and the examinees do not guess. For the multidimensional three-parameter logistic model, the parameters estimated are the ***a***-, *d*-, *and c*-parameters. In this model, it is assumed that the examinees guess when they do not know the correct answer. It should be noted that in the case of multidimensional models the ***a***-parameter is a vector where each element is the discrimination parameter for each dimension.

Since the terms in the exponent are additive, an examinee who has a low ability on one dimension can be compensated for by a high ability on a second dimension

(Ackerman, 1994). In MIRM, the probability of correct response to an item on two or more dimensions is represented by the item characteristic surface (ICS). Figure 4 shows the probability of correct response to items 1 and 2 given the examinees composite abilities on two dimensions.

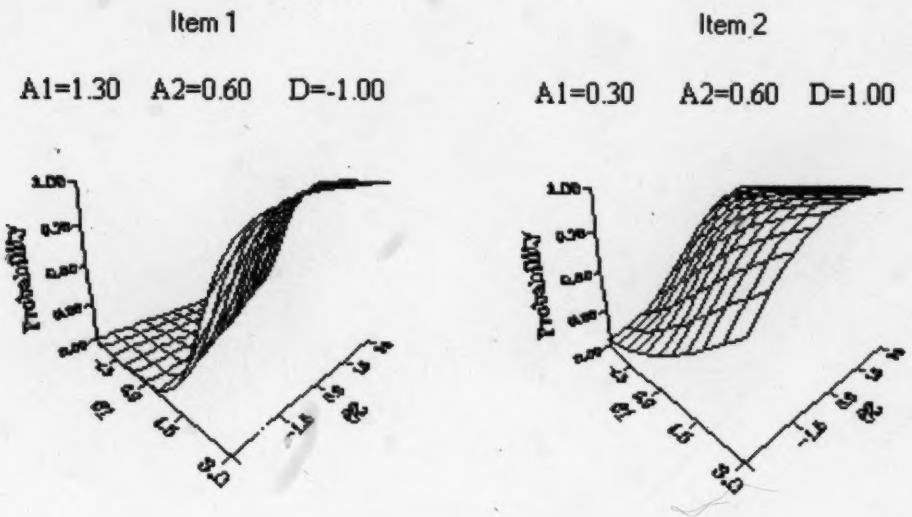


Figure 4. Item Characteristics Surface.

However, interpreting information presented by the ICS is difficult. A contour plot of the ICS provides a better representation of the item. Each contour represents the probability of an examinee with a given composite ability answering the item correctly. For example, examinees with the same probability of answering the items correctly lie on the same contour line. For the compensatory model, the contours are parallel and equally spaced across the response surface (Ackerman, Gierl, & Walker, 2003). The higher the discrimination of the item, the closer the contour lines and the direction of maximum slope will always be perpendicular to those lines. As shown in Figure 5, the contour

lines are closer to each other for item 1 because item 1 is more discriminating than item 2.

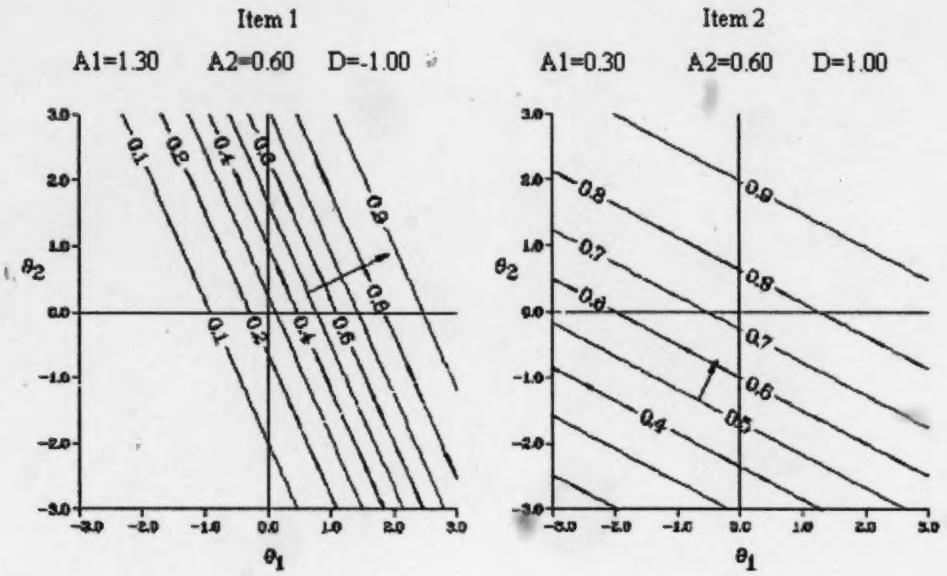


Figure 5. Contour Plot of Item Characteristic Surface.

The limitation of the contour plots is that only one item can be depicted at a time. Reckase and McKinley (1991) developed the vector plot where all the items of a test can be presented in one plot. An example is provided in Figure 6. The items are depicted as vectors in a Cartesian coordinate system where the length of the vector corresponds to the amount of multidimensional discrimination $MDISC_i$ (Ackerman, 1994; Ackerman, Gierl, & Walker, 2003).

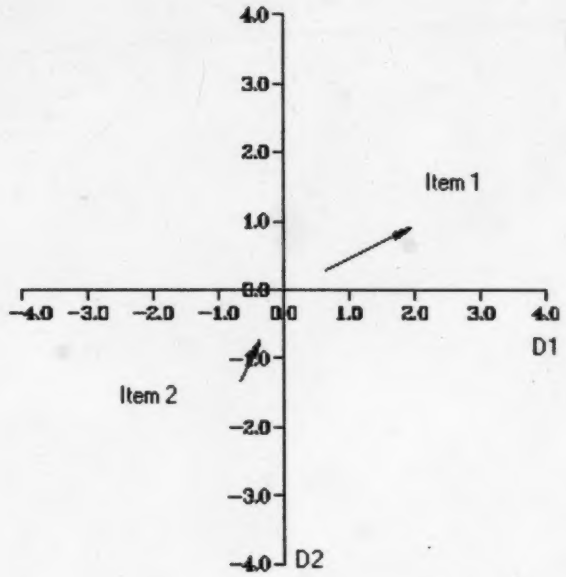


Figure 6. Vector Plot of Relatively Low and High Discriminating Items.

In the case of two dimensions, the $MDISC_i$ is given by:

$$MDISC_I = \sqrt{a_{i1}^2 + a_{i2}^2},$$

where a_{i1} and a_{i2} are the discrimination parameters of dimensions 1 and 2, respectively.

All item vectors lie on a line that passes through the origin creating an angle with the positive θ_I (Ackerman, 1994). The origin is the population multidimensional ability level mean. The angle with the positive θ_I specifies the direction of the vector where the item has maximum discrimination. This direction is referred to as the direction of best measurement (Ackerman, 1994; Stout *et al.*, 1996). The angular direction is given by (Reckase & McKinley, 1991):

$$\alpha_i = \arccos \frac{a_{i1}}{MDISC_I}$$

The vector originates at and is graphed orthogonal to the $p = 0.50$ equiprobability contour (Ackerman, 1994). Since the *a*-parameters are restricted to be positive, the vectors are located only in the first and the third quadrants (Ackerman, 1996; Ackerman, Gierl, & Walker, in press). Figure 6 depicts a relatively high discriminating item (Item 1) and a relatively low discriminating item (Item 2). The length of the vector of item 1 clearly indicates that this item is more discriminating than item 2.

The difficulty of the item *D* corresponds to the location of the vector in space (Ackerman, Gierl, & Walker, 2003) and is represented by the signed distance from the origin to the $p = 0.50$ equiprobability contour. The multidimensional difficulty is given by (Reckase, 1985):

$$D = \frac{-d_i}{MDISC_i},$$

where

d_i is a scalar related to the difficulty of item *i*, and

$MDISC_i$ is the multidimensional discrimination parameter for item *i*.

Items that have positive *D* are relatively difficult and lie in the first quadrant whereas items that have negative *D* are relatively easy and lie in the third quadrant. Figure 6 depicts both an easy item and a difficult item. Item 1 lies in the first quadrant, thus is more difficult ($D = 0.70$) than item 2 that lies in the third quadrant ($D = -1.49$).

Assumptions

The four main assumptions of the compensatory MIRM model are:

1. Examinees abilities are assumed to be randomly and independently distributed with an *m*-variate normal distribution.

2. Each dimensional component has a mean of 0 and variance of 1 in the population in which the items are calibrated.
3. The compensatory nature of the model assumes that examinees with low ability on one dimension can compensate by having a high ability on a second dimension.
4. After controlling for abilities, the examinees responses to the test items are independent.

Score Estimation

Expected A Posteriori

In the case of multidimensional item response models, the EAP is used to estimate an *m x 1* vector of abilities for each examinee. In the case of two dimensions, two abilities are estimated -- one for each of the dimensions. The elements of the ability vector for each examinee are approximated by Bayes mean estimators computed marginally over the joint posterior ability distributions (Luecht & Miller, 1992). That is,

$$\hat{\theta}_{j1} = \sum_{k=1}^{Q} t_k f(t_k, t), k = 1, ..., Q,$$

where

t_k is the weight or probability density at the k^{th} quadrature point along the θ_1 axis,

Q is the number of quadrature points on the θ_1 axis,

$\hat{\theta}_{j1}$ is the estimated ability for examinee *j* on dimension 1.

and

$$\hat{\theta}_{j2} = \sum_{l=1}^{R} t_l f(t, t_l), l = 1, ..., R,$$

where

t_l is the weight or probability density at the l quadrature point along the θ_2 axis,

R is the number of quadrature points on the θ_2 axis, and

$\hat{\theta}_{j2}$ is the estimated ability for examinee j on dimension 2.

The joint posterior distribution is given by:

$$f(t_k,t_l) = \frac{L(U/t_k,t_l)G(t_k,t_l)}{\sum_{k=1}^{Q}\sum_{l=1}^{R} L(U/t_k,t_l)G(t_k,t_l)},$$

where

$L(U|t_k,t_l)$ is the likelihood taken across the items and is given by:

$$L(U|t_k,t_l) = \prod_{i=1}^{N} P(t_k,t_l)^{u_i} Q(t_k,t_l)^{1-u_i}, \text{ and}$$

$G(t_k,t_l)$ is a bivariate normal prior distribution.

Advantages and Limitations of Scoring Using MIRMs

The advantage of obtaining score estimates using the MIRMs is that it results in appropriate ability estimates when the data are multidimensional. The ability of the examinee can be estimated on each dimension. Further, ability parameter estimates are not test-dependent and the error of measurement is not the same across the score scale on each dimension.

However, the MIRMs have several disadvantages. First, the MIRMs require large sample sizes. This factor limits the use of MIRMs to large-scale assessments.

Second, the MIRMs score estimation is complex and requires advanced statistical knowledge and high-speed computers to conduct the estimation and interpret the results.

Review of Previous Research Studies

Various studies have been conducted in which different scoring models using selected and constructed response items were compared. Since this study compares four scoring methods using dichotomously scored items, the following literature review is restricted to the studies that only considered selected response items. The literature review revealed seven relevant studies. Four of these studies involved UIRMs; the remaining three involved MIRMs.

Previous Research: UIRM

Anderson (1999) compared the number-right and 3-parameter item response scoring models for the 1996 British Columbia diploma examination for Grade 12 Mathematics. The mathematics test included 50 multiple-choice items that Anderson used as an item bank. Two samples of 25 items each were drawn where one sample consisted of items with even numbers and the other sample consisted of items with odd numbers. The correlation between the scores yielded by the two methods and the Root Mean Square Difference (RMSD) were used as measures of similar ranking and difference, respectively. The RMSD was calculated for each set of scores using the percent correct on the 50 items as the domain score and on the 25 item tests as the estimated scores.

Anderson found that the scores yielded by the number-right and the 3-parameter item response models were similar in value. For the two tests, the correlation between scores yielded by the number-right and the 3-parameter model were very high ($r = 0.971$

for test 1 and $r = 0.977$ for test 2). The root mean square differences between scores yielded by the two scoring methods and the domain scores were low (RMSD = 0.060 for test 1 and 0.060 for test 2 using number-right scoring and RMSD= 0.067 for test 1 and 0.065 for test 2 using the 3-parameter scoring model). The two methods also yielded scores that classified students similarly. Based on these results, Anderson concluded that the use of complex item response scoring models may not be warranted.

Anderson noted two limitations of his study. First, the odd-even number split that was used to create the two tests was arbitrary; therefore, the resulting two tests were not parallel. Second, the results of the study should be interpreted cautiously because the item pool was small (contained only 50 items) and only two samples of 25 items each were drawn.

Fan (1998) compared ability estimates yielded by the classical test score model and the one-, two-, and three-parameter item response models using the 1992 Texas Assessment of Academic Skills (TAAS) administered to grade 11 students. The TAAS is a large-scale criterion-referenced test designed to assess the mastery of school instructional objectives and consists of reading, writing, and math tests. Only the reading and mathematics tests were used in his study. The reading and mathematic tests consisted of 48 and 60 multiple-choice items, respectively. For each test, 20 samples of 1000 examinees were randomly selected. The average correlations between ability estimates and obtained scores across the 20 samples were used to assess the comparability of person statistics. The average correlations between the classical test score model and the one-, two-, and three-parameter item response models were obtained

by transforming the correlations to Fisher *zs*, computing the average, and then transforming the average back to Pearson's correlations.

Fan found that the ability estimates resulting from the two scoring frameworks were quite comparable. The average correlations of the ability estimates between the classical test score model and the one-, two-, and three-parameter item response models were very high and ranged from 0.966 to 0.997.

Fan acknowledged two limitations in his study that may undermine the validity of his findings. First, there was a strong ceiling effect suggesting that many items were easy. Many items tended to be less variable in terms of item difficulty and discrimination. This is typical of criterion-referenced tests where most items are answered by 80% to 100% of the students. The effects of the low variability of the item parameters on the results are not known. Second, the item pool was not large enough for items with varying characteristics to be sampled. Therefore, different samples of items may have produced different results.

Ndalichako and Rogers (1997) compared number-right, one-, two-, three-parameter item response models, and finite–state scoring models. Since the finite-state scoring model was not included in the present study, only the results relevant to the number-right and the one-, two-, and three-parameters item response scoring models are discussed here. A sample of 1230 examinees writing the multiple-choice section of a school-leaving reading comprehension test was used for the comparisons. The test consisted of 70 four-option multiple-choice items. The correlation between the ability estimates yielded by the different scoring models and the mean absolute difference (MAD) between the transformed scores were used as measures of score comparability.

Ndalichako and Rogers (1997) found that the correlation between the ability estimates yielded by the scoring models were nearly perfect. The correlations ranged from 0.977 to 0.994. Rogers and Ndalichako (2000) reported similar findings where the correlation between the number-right, one-, two-, and three-parameter item response models ranged from 0.984 to 0.999. The correlations values reported by Ndalichako and Rogers (1997) and Rogers and Ndalichako (2000) were also similar to the correlations values reported by Fan (1998).

Furthermore, the MAD among the pairs of transformed scores ranged from 0.77 to 1.53. The closest agreement was between the one- and two-parameter models (0.77), and the two- and three-parameter models (0.79). Compared to the one-, two-, and three-parameters scoring models, the number-right agreed more closely with the 3-parameter model (1.21) than with the one-parameter model (1.53) or the two-parameter model (1.36).

Tomkowicz and Rogers (in press) compared examinees ability estimates yielded by the number-correct, one-, two-, three-parameter item response models, and the nominal response model in the presence of items susceptible to testwiseness. Since the nominal response model was not included in the present study, only the results relevant to the number-right and the one-, two-, and three-parameters item response scoring models are discussed here. A random sample of 4000 grade 12 students who wrote the Social Studies and the Chemistry diploma examinations was used for score comparison. The Social Studies test consisted of 70 items and the Chemistry test consisted of 44 items. Two panels of experts separated the multiple-choice items of the two tests into items susceptible to testwiseness strategies and items not susceptible to testwiseness

strategies. The students were also separated into high, middle, and low ability groups. The five scoring methods were compared in terms of similar ranking of examinees, agreement of values between the ability estimates and the number correct, and the proportion of examinees that received ability estimates that were greater than or less than one standard error of measurement of the number-right estimate.

Tomkowicz and Rogers found that the two- and three-parameter response models yielded ability estimates that were different from the number correct and the one-parameter model across the three ability groups. This difference was greater for Chemistry than for Social Studies. For Chemistry, the correlation between the number-right and the one-parameter model ranged from 0.99 to 1.00 and the root mean squared difference ranged from 0.55 to 1.09 across the three ability groups. In contrast, the correlation between number-right, two- and three-parameter models ranged from 0.92 to 0.97 and the root mean squared difference ranged from 1.48 to 2.39 across the three ability groups. For Social Studies, the correlation between the number-right and the one-parameter model ranged also from 0.99 to 1.00 and the root mean squared difference ranged from 0.81 to 1.60 across the three ability groups. In contrast, the correlation between number-right, two-, and three-parameter models ranged from 0.94 to 0.97 and the root mean squared difference ranged from 1.45 to 2.00 across the three ability groups.

Furthermore, with the exception of high ability group in Social Studies, the difference was more pronounced for the subtests that contained items susceptible to testwiseness. The difference was the greatest for students with low ability. For Chemistry, the correlation between the number-right and the one-parameter model was 0.99 for the three groups and the root mean squared difference ranged from 0.59 to 0.99

across the three ability groups. In contrast, the correlation between number-right, two- and three-parameter item response models ranged from 0.85 to 0.93 and the root mean squared difference ranged from 2.26 to 3.54 across the three ability groups. For Social Studies, the correlation between the number-right and the one-parameter model ranged from 0.99 to 1.00 for the three groups and the root mean squared difference ranged from 0.69 to 1.33 across the three ability groups. In contrast, the correlation between number-right, two-, and three-parameter item response models ranged from 0.94 to 0.97 and the root mean squared difference ranged from 1.67 to 1.98 across the three ability groups.

Previous Research: MIRM

Luecht (2003) compared four scoring methods for reporting sub-scores for diagnostic purposes using the multidimensional compensatory three-parameter item response model. A simulated dichotomous data set composed of a sample of 2000 examinees and 74 items was used. The item parameters were derived from one of the multiple sections of a certification test that provided fail/pass decisions. The section covered four professional competency areas. The item parameters that were originally calibrated by the three-parameter model were used as known parameters. A four factor (corresponding to the four competency areas) oblique simple structure with correlation of 0.50 between the pairs of the four abilities was set to produce the simulated data. The four scoring methods were: standardized number correct scores Z (X); Bayes mean (EAP) unidimensional total-test calibration UIRT (T); Bayes mode (MAP) based on separate unidimensional calibrations of items for the separate abilities UIRT (S); and MAP based on a multidimensional total test calibration (MIRT), with one factor representing each of two abilities.

Luecht reported that the UIRT (S) and the MIRT methods provided the most accurate estimates whereas the UIRT (T) had the largest average standard error followed by the Z (X) method. The average standard errors of UIRT (S) and MIRT ranged from 0.15 to 0.57 and from 0.20 to 0.66, respectively, across the four competency areas. In contrast, the average standard errors of UIRT (T) and Z (X) ranged from 0.43 to 1.61 and from 0.42 to 0.78, respectively, across the four competency areas. Furthermore, Luecht compared estimated score profiles of two examinees that had the same total score, but heterogeneous subtest scores profiles across the four competency areas, to their true profiles (the true subtest scores profiles were known because the multidimensional generating parameters were known). He found that the UIRT (T) and UIRT (S) produced subtest scores that were more consistent with the true score profiles than Z (X) and MIRT. Luecht concluded that since UIRT (S) and MIRT were the most precise. He recommended the use of UIRT (S) because MIRT estimation is much more complex.

Luecht acknowledged that since the comparisons between the scoring methods were based on a single simulation study, the results were not conclusive. He also acknowledged the shortcoming of the TESTFACT 4.0 program (Bock *et al.*, 2003) used to calibrate the items and produce subtest scores in the multidimensional case MIRT. Luecht noted that the "MIRT factors can be misinterpreted if the items are not specifically constrained to load on particular factors. Unfortunately, TESTFACT does not provide the capability to implement such constraints" (Luecht, 2003, p. 11).

However, there are two additional concerns that must also be addressed. First, several estimation procedures were used to obtain ability estimates instead of one estimation method. The differences noted between the four scoring methods may be in

part due to the different estimation methods. Second, since simple structure was used, we do not know the effect of complex structure on the differences between the four scoring methods.

Recently, Tate (2004) examined whether reporting multiple sub-scores offered useful diagnostic information and the extent of total score degradation as a function of dimensionality and the level of correlation among the abilities. He also examined whether *maximum likelihood estimation* (MLE) or *expected a posteriori* (EAP) provided the best estimation procedure, after rescaling, for sub-scores and total test scores. The multidimensional Rasch model was used where, for each dimension, the item difficulties were evenly distributed at five values between –1.0 and +1.0. The test dimensionality, the level of correlation among the dimensions, the number of items in the subtests, and the estimation procedures were varied. The values for the number of dimensions were 2, 3, 4, and 5. The values of the correlations between the dimensions were 0.0, 0.2, 0.4, 0.6, 0.8, 0.9, and 0.95. The number of items in the subtests was 12, 15, 20, and 30. A sample of 1000 examinees responding to a 60-item test was generated. The items in the test were dichotomously scored. The distributions of all abilities were assumed to have a mean of 0.0 and a variance of 1.0. The mean error variance was used as a measure of accuracy of the two estimation methods and the mean absolute deviation was used to evaluate the subtest score differences.

Tate found that the choice between the two estimation methods, after rescaling, depended on the intended uses of the sub-scores to make relative decisions (norm-referenced) or absolute decisions (criterion-referenced). For correlation levels of 0.6 or above the MLE was superior to the EAP for relative decisions regardless of the number

of items in the subtests. The difference between MLE and EAP was small for low correlation and increased as the correlation levels increased. In contrast, the EAP was superior to the MLE for absolute decisions when the number of items in the subtests was small. The differences between the MLE and EAP estimates in terms of error variance were relatively small when the number of items in the subtests was large (30 items) or when the correlation levels were between 0.4 and 0.6 for any number of items in the subtest. However, the MLE was superior for low correlation and the EAP was superior for high correlation when the number of items in the subtest was small (12 items) and the correlation levels were extreme. For the total test score, the MLE estimates were superior to the EAP estimates for all combinations of dimensionality and correlation levels. The differences in mean error variance between the MLE and the EAP were small for correlations of 0.6 and above or for dimensionality of 2. However, the difference in mean error variance between MLE and EAP was relatively larger for low correlation and high dimensionality.

Tate acknowledged several limitations in his study. First, since the data was based on the one-parameter item response model, the results may differ to some degree for the two- and three-parameter item response models. Similarly, a generalization to other situations cannot be made for tests having unsymmetrical structure; tests with different numbers of items in the subtests and different numbers of items in the total test; tests with different assumed levels and distributions of item difficulties; and tests where the correlations are different for different pairs of abilities. Second, the scores yielded by the MLE and EAP cannot be used for follow up research to describe the relationship among the sub-scores because the correlations among the MLE subtest scores are

negatively biased whereas the correlation among the EAP sub-scores are positively biased.

However, other limitations must also be addressed. First, a simple structure was assumed. It is not known what the effects on sub-scores and total scores performance would be if a complex structure was used. Second, each one of the abilities distributed along the dimensions was assumed to have a mean of 0.0 and variance of 1.0. The effect on the performance of sub-scores and total scores if the dimension abilities have different means is not known. Finally, Tate noted that a comparison between the actual error variances based on the known true abilities and the reported error variances for the EAP estimates of the total score obtained from the ConQuest program (Wu *et al.*, 1998) indicated "an actual precision that was appreciably better than the apparent EAP values" (Tate, 2004, p. 92). These results raised doubts about the accuracy of the reported EAP error variances obtained from the ConQuest program for at least some conditions when estimating total scores in the multidimensional case. A detailed examination of the accuracy of the reported EAP error variances was not provided because it was beyond the scope of his study. Therefore, there is a doubt about the accuracy of the estimated error variance for EAP that may have affected the results.

Previous Research: Classification

Walker and Beretvas (2003) examined the effect of using the unidimensional 3-parameter model and the 3-parameter multidimensional models on classifying examinees into four levels of proficiency. A subset of data obtained from the fourth and seventh

grade 1998 Mathematics administration of the *Washington Assessment of Student Learning* (WASL) was used. The data included 63, 533 fourth graders and 65, 279 seventh graders. The open-ended items of the Mathematics test were chosen for the analysis because these items were hypothesized to be two-dimensional where the first dimension was considered to measure general mathematics ability and the second dimension was considered as measuring mathematical communication. The NOHARM II and 2D-EAP programs were used to estimate the unidimensional and the multidimensional item parameters and the ability parameters, respectively. The pseudo-guessing parameters were estimated using MULTILOG IV. Walker and Beretvas found that examinees that were proficient in mathematical communication tended, on average, to be classified by the unidimensional model into higher proficiency levels compared to the multidimensional model. Similarly, examinees whose abilities were low in mathematical communication were more likely to be classified into lower mathematical ability levels when the unidimensional model was used compared to the multidimensional model.

However, the authors acknowledged several limitations. Although the open-ended items were polytomously-scored, the data was dichotomized to meet the requirements for the available multidimensional models and ability estimation programs. Further, the correlation between the two dimensions was moderate to high with a correlation of 0.61 between the two dimensions reported for the fourth grade and a correlation of 0.81 reported for the seventh grade. Finally, the comparisons were made based on the assumption that the data was two-dimensional, but the true dimensionality of the data was not known.

Summary of the Limitations of Previous Research

Taken together, the studies in the existing literature have a number of important limitations. For example, Anderson (1999) and Fan (1998) used small pools of items. Anderson also used non-parallel tests whereas Fan used items with low variability of the difficulty and discrimination parameters. The more relevant studies of Luecht (2003) and Tate (2004) shared the limitation of restricting their studies to tests with simple structure. Luecht also used several estimation procedures, did not replicate the results, and the multidimensional item parameters were obtained by TESTFACT using exploratory factor analysis. Tate used the one-parameter model, restricted the ability distribution for each dimension to mean of 0.0 and variance of 1.0, and expressed doubt about the accuracy of the reported EAP error variances obtained from the ConQuest program for some conditions. Walker and Beretvas (2003) dichotomized the polytomously-scored items, the correlation between the two dimensions was high, and the true dimensionality of the data was unknown.

Summary

The literature review of the articles related to this study indicated that the literature on comparing different scoring methods when the data are multidimensional is scarce. Seven articles relevant to this study were found, only three of which dealt with multidimensional models. Given the limited research in this area, one of the aims of this study was to fill in the gap where the accuracy of these scoring methods had not been thoroughly investigated when the data were known to be multidimensional and when different conditions were present.

This study attempted to overcome many of the limitations of the previous research by including the following characteristics: more variability in the items difficulty and discrimination parameters; using one score estimation procedure; confirmatory approach; simple versus complex structure; one hundred replications; less restrictive model than the one-parameter model; and variable ability distributions in terms of varying the mean on the second dimension. In addition, various correlations between the examinees' abilities were included to investigate its effects on the accuracy of the total score and the sub-scores. Finally, the dimensionality of the data was known.

CHAPTER 3: METHOD

Overview

The purpose of the present study was to assess the degree to which the CTM, UIRM (T), and UIRM (S) scoring models were able to recover the scores obtained using the MIRM scoring model when the dimensionality of the test was known to be two. Ideally, the true ability values would be used for comparison. However, the lack of appropriate software to achieve this purpose prevented the use of the true ability values. Consequently, the estimated abilities from the MIRM scoring model were used as the base for the comparisons made.

To address this purpose, a simulation study was conducted. The factors considered included:

1. correlation between examinees abilities on the two dimensions: 0.0, 0.3, 0.6, and 0.9;
2. mean differences on the two dimensions: equal (0,0) and unequal (0, -1);
3. factor complexity: simple and complex; and
4. type of score reported: total score and subtest score.

The methods, procedures, and computer programs used to address these objectives are discussed in this chapter.

Method

Data

This study was conducted using simulation. Each simulated data set consisted of 2000 examinees responding to a 40-item test. The sample of 2000 examinees was chosen based on the finding that the error in estimating the multidimensional item

parameters greatly increased when the sample size was less than 2000 examinees (M. D. Reckase, personal communication, July 26, 2005). The samples of respondents were selected randomly from the conceptual population corresponding to the randomly generated samples of respondents to allow generalization of the results. The 40-item test was designed to measure two underlying dimensions where items 1 to 20 measured the first dimension (D1) and items 21 to 40 measured the second dimension (D2). The equal number of items in the two subtests created a test that measured both dimensions equally in terms of the number of items. In addition, the variance of the items in the subtests was similar giving equal weights to the two subtests. The generated responses to these items were dichotomously scored.

Multidimensional Item Parameters

To make the simulation more realistic, the multidimensional item parameters used to generate the data were based on the item parameters obtained from the October 1992 administration of the Law School Admission Test (LSAT) (*Law School Admission Council*, 1999). The LSAT consists of three subtests that appear to measure analytical reasoning (AR), logical reasoning (LR), and reading comprehension (RC). Douglas *et al.* (1999) found that, in addition to several weaker secondary dimensions, the LSAT appears to have two dominant dimensions -- the first dimension corresponds to the AR section and the second dimension corresponds to the combined LR and RC sections. The total number of items in the October 1992 LSAT was 102 items where 78 items measured the combined LR and RC dimension and 24 items measured the AR dimension. The *a*-parameters in this dissertation were set greater than the mean of the LSAT *a*-parameters. The *d*-parameter range in the LSAT was not balanced (-1.0 to

1.95); a more balanced *d*- parameter range was used in this study. Table 1 presents the item parameters means and standard deviations of the LSAT.

Table 1

LSAT Item Parameters for October 1992 Administration.

		Statistics			
Parameters	N (Items)	Min	Max	Mean	SD
d	102	-1.00	1.95	0.25	0.61
a_1	78	0.21	0.78	0.47	0.12
a_2	24	0.29	0.77	0.55	0.15

Since the 2-parameter, 2-dimensions, compensatory model was used in this study to generate the data, the *c*-parameter was set to zero for all items. The mean *d*-parameter for each of the two dimensions was kept constant to eliminate the confounding effect of differing mean item difficulty on the ability estimates. However, the *d*-parameter within each of the dimensions ranged from –1.0 to +1.0 with a mean of 0.0 and a standard deviation of 0.64 (see Table 2).

The item discrimination parameters were set for D1 and D2 to create two conditions: simple structure and complex structure. For the simple structure, the item discrimination parameters a_1 for items measuring D1 and a_2 for items measuring D2 were kept constant at 0.60 ($M = 0.60$, $SD = 0.00$). This ensured that the differences in ability estimates on the two dimensions were not due to one dimension having more discriminating items than the other dimension. However, the discrimination parameters a_2 for items measuring D1 and the a_1 for items measuring D2 were varied to create

different angles with respect to D1. The range of these items discrimination parameters was between 0.00 and 0.21 ($M = 0.10$, $SD = 0.06$) (see Table2).

For the complex structure, the item discrimination parameters a_1 for items measuring D1 and a_2 for items measuring D2 was kept constant at 0.60 ($M = 0.60$, $SD = 0.00$). However, the discrimination parameters a_2 for items measuring D1 and the a_1 for items measuring D2 were varied. The range of these items discrimination parameters was from 0.28 to 0.58 ($M = 0.42$, $SD = 0.09$) (see Table 2).

Table 2

Item Parameters for Simple and Complex Structure

	Simple Structure				Complex Structure			
Item	a_1	a_2	d	c	a_1	a_2	d	c
1	0.60	0.21	-1.00	0.00	0.60	0.58	-1.00	0.00
2	0.60	0.20	-0.90	0.00	0.60	0.56	-0.90	0.00
3	0.60	0.18	-0.80	0.00	0.60	0.54	-0.80	0.00
4	0.60	0.17	-0.70	0.00	0.60	0.52	-0.70	0.00
5	0.60	0.16	-0.60	0.00	0.60	0.50	-0.60	0.00
6	0.60	0.15	-0.50	0.00	0.60	0.49	-0.50	0.00
7	0.60	0.14	-0.40	0.00	0.60	0.47	-0.40	0.00
8	0.60	0.13	-0.30	0.00	0.60	0.45	-0.30	0.00
9	0.60	0.12	-0.20	0.00	0.60	0.44	-0.20	0.00
10	0.60	0.11	-0.10	0.00	0.60	0.42	-0.10	0.00
11	0.60	0.10	0.10	0.00	0.60	0.40	0.10	0.00
12	0.60	0.08	0.20	0.00	0.60	0.39	0.20	0.00
13	0.60	0.07	0.30	0.00	0.60	0.37	0.30	0.00
14	0.60	0.06	0.40	0.00	0.60	0.36	0.40	0.00
15	0.60	0.05	0.50	0.00	0.60	0.35	0.50	0.00
16	0.60	0.04	0.60	0.00	0.60	0.33	0.60	0.00
17	0.60	0.03	0.70	0.00	0.60	0.32	0.70	0.00
18	0.60	0.02	0.80	0.00	0.60	0.31	0.80	0.00
19	0.60	0.01	0.90	0.00	0.60	0.29	0.90	0.00
20	0.60	0.00	1.00	0.00	0.60	0.28	1.00	0.00
21	0.21	0.60	-1.00	0.00	0.58	0.60	-1.00	0.00
22	0.20	0.60	-0.90	0.00	0.56	0.60	-0.90	0.00
23	0.18	0.60	-0.80	0.00	0.54	0.60	-0.80	0.00
24	0.17	0.60	-0.70	0.00	0.52	0.60	-0.70	0.00
25	0.16	0.60	-0.60	0.00	0.50	0.60	-0.60	0.00
26	0.15	0.60	-0.50	0.00	0.49	0.60	-0.50	0.00
27	0.14	0.60	-0.40	0.00	0.47	0.60	-0.40	0.00
28	0.13	0.60	-0.30	0.00	0.45	0.60	-0.30	0.00
29	0.12	0.60	-0.20	0.00	0.44	0.60	-0.20	0.00
30	0.11	0.60	-0.10	0.00	0.42	0.60	-0.10	0.00
31	0.10	0.60	0.10	0.00	0.40	0.60	0.10	0.00
32	0.08	0.60	0.20	0.00	0.39	0.60	0.20	0.00
33	0.07	0.60	0.30	0.00	0.37	0.60	0.30	0.00
34	0.06	0.60	0.40	0.00	0.36	0.60	0.40	0.00
35	0.05	0.60	0.50	0.00	0.35	0.60	0.50	0.00
36	0.04	0.60	0.60	0.00	0.33	0.60	0.60	0.00
37	0.03	0.60	0.70	0.00	0.32	0.60	0.70	0.00
38	0.02	0.60	0.80	0.00	0.31	0.60	0.80	0.00
39	0.01	0.60	0.90	0.00	0.29	0.60	0.90	0.00
40	0.00	0.60	1.00	0.00	0.28	0.60	1.00	0.00

Simple Versus Complex Structure

To create simple structure and complex structure, the angular directions of the item vectors relative to the positive D1 axis were varied. The angular direction relative to the positive D1 axis is given by:

$$\alpha_{rad} = \cos^{-1} \frac{a_1}{MDISC_i},$$

where

α_{rad} is the angle between the item *i* and D1 in radian units,

cos^{-1} is the arccosine of the angle,

a_1 is the item discrimination for D1, and

$MDISC_i$ is the multidimensional discrimination parameter.

The angular direction formula yields numbers in radian units. To convert angles measured in radian units to degrees, the following formula was used:

$$\alpha_{deg} = \frac{180^{\circ}}{\pi} \alpha_{rad}.$$

For the simple structure, the first 20 items measured primarily D1 and the last 20 items measured primarily D2. The angular direction of the items relative to the positive D1 axis ranged from 0° degrees to 19° degrees for D1, with an increment of 1°degree between the items ($M = 9.50^{\circ}$ and $SD = 5.92^{\circ}$). In contrast, the angular direction of the items relative to the positive D1 axis ranged from 71° degrees to 90° degrees for D2 with an increment of 1° degree between the items ($M = 80.50^{\circ}$ and $SD = 5.92^{\circ}$) (Luecht & Miller, 1992).

For the complex structure, the 40 items in this test measured both D1 and D2 to a greater extent and in varying degrees. The angular direction of the items relative to the

positive D1 axis ranged from 25° degrees to 44° degrees for D1, with an increment of 1° degree between the items (M = 34.50° and SD = 5.92°). In contrast, the angular direction of the items relative to the positive D1 axis ranged from 46° degrees to 65° degrees for D2 with an increment of 1° degree between the items (M = 55.50° and SD = 5.92°). The overall mean and standard deviation of the degrees for the 40 items was 45.00° and 12.13° respectively.

Mean of Ability Distributions

To study the effect of varying the mean of the ability distribution on the ability estimates yielded by the CTM, UIRM (T), UIRM (S), and MIRM, the ability distribution corresponding to the first dimension was kept constant with a mean of 0.0. In contrast, the mean of the ability distribution corresponding to the second dimension was varied. The two mean values for the ability distribution on the second dimension were 0.0 and -1.0. However, the variance of the ability distribution corresponding to the first and the second dimension remained constant at 1.0. Figure 7 depicts ability distributions with mean vector (0.0, 0.0). As shown in Figure7, the mean of the ability distribution corresponding to each of the two dimensions was kept at 0.0.

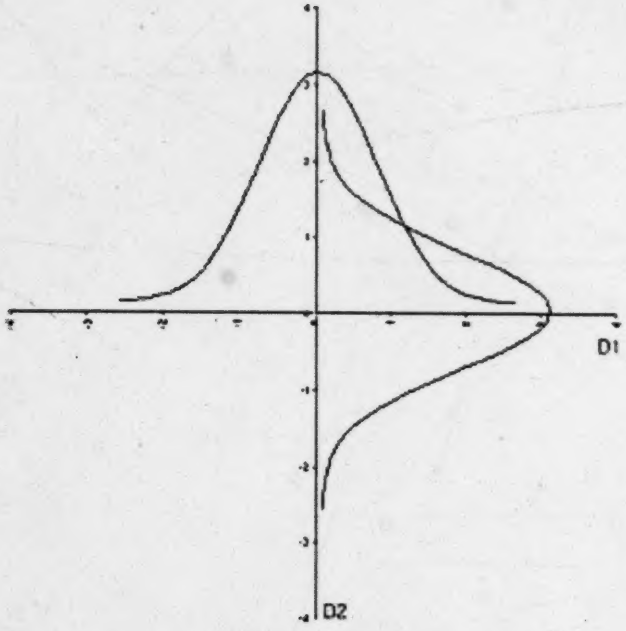


Figure 7. Ability Distributions with Mean Vectors (0, 0).

Figure 8 depicts ability distributions with mean vector (0.0, -1.0). In contrast to the ability distributions in Figure 7, the mean of the ability distribution corresponding to the first dimension in Figure 8 was kept at 0.0 whereas the mean of the ability distribution corresponding to the second dimension was shifted to –1.0.

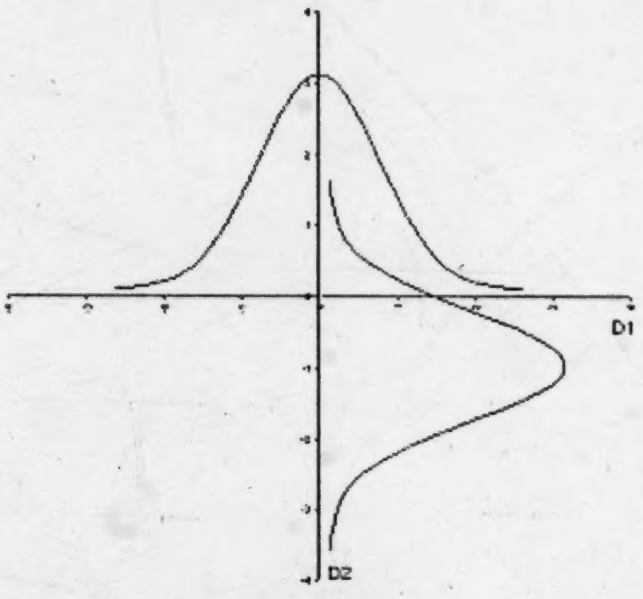


Figure 8. Ability Distributions with Mean Vectors (0, -1).

Correlation Between Abilities

The correlation between pairs of ability estimates was also varied to investigate if the CTM, UIRM (T), and UIRM (S) ranked the examinees in the same order as the MIRM as a function of the correlation between abilities. The four correlation levels were 0.0, 0.3, 0.6, and 0.9. These correlation values were selected to span the range from no relationship (r = 0.0) to very strong relationship between the two sets of abilities (r = 0.9).

Design

The design of this study can be viewed as a 4 (correlation) x 2 (means) x 2 (structure) x 2 (scores) fully crossed design. The combination of the two means of the ability distribution on the second dimension and the four correlation values resulted in 8

conditions. Since these conditions were considered under simple structure and complex structure, a total of 16 conditions resulted. Furthermore, total scores and subtest scores were computed for each of these conditions, which resulted in a total of 32 conditions.

Procedure

Data Generation

For each condition, the program Multidimensional Simulation MULTISIM (Stout, 1990) was used to generate 100 data sets. The purpose of generating 100 data sets was to replicate the results a large number of times using different samples in order to assert that the observed difference between ability estimates was not due to chance. MULTISIM is a FORTRAN program designed to simulate multidimensional data up to four dimensions. For each condition, the multidimensional item parameters (a_1, a_2, d, and c), the mean and the variance of the ability distribution for each dimension, the correlation between the two dimensions, the number of examinees, and the number of replications were specified in the input file.

The number of data sets and the random number seed were also specified in the input files. The random number seed was different for each condition defined by the correlation value, mean, and variance. For each condition, the MULTISIM program created an output file containing all of the 100 data sets required. The data were then parsed into separate data files using the Statistical Package for the Social Sciences (SPSS) program syntax. The four scoring methods were then applied. This ensured that a common sample was used for the comparisons among scores.

Classical Test Model Scoring Model

Total Test Score CTM (T)

The number of correct responses was summed for each examinee to obtain the total test score. Since the BILOG 3.11 (Mislevy & Bock, 1997) gives, as part of its output, the total number of correct responses, the total score for the CTM was obtained using BILOG 3.11. The Binary Logistic models software (BILOG 3.11) is a Windows program used to estimate item and ability parameters using the unidimensional binary logistic models. The output of the BILOG 3.11 program consists of three phases. Phase 1 provides item statistics based on the CTM. Phase 2 provides items parameters estimates based on the UIRM. Phase 3 provides the total test score and the examinees' ability estimates.

Subtest scores CTM (S)

The number of correct responses for the first 20 items measuring dimension 1 was summed to obtain subtest score 1. Similarly, the number of correct responses for the last 20 items measuring dimension 2 was summed to obtain subtest score 2. When two subtests were specified in the input file of the BILOG 3.11 program, the program produced ability estimates for each subtest along with total number correct for each subtest.

Ability Estimation

As previously mentioned (p. 25), the *expected a posteriori* (EAP) was used to estimate the ability of examinees for the UIRM and the MIRM scoring. The use of one estimation method maintained consistency of ability estimation across the UIRMs and MIRMs and eliminated any difference in abilities that may be due to using different

estimation procedures. The advantages of EAP over the other estimation methods were previously discussed in Chapter 2 (p. 26).

Unidimensional Item Response Scoring Models

Total Test Score UIRM (T)

The BILOG 3.11 (Mislevy & Bock, 1997) program was used to estimate the ability of the examinees on the total number of items in the test. The test was specified as one test in the BILOG 3.11 input file; therefore, the BILOG 3.11 performs single calibration of the items. One ability estimate for each examinee was produced as part of the BILOG 3.11 output and was used as the total score ability estimate.

Subtest scores UIRM (S)

As in UIRM (T), the BILOG 3.11 program was used to estimate the ability of the examinees. However, the test was specified in the input file as comprising of two subtests. Items 1- 20 were specified as subtest 1 and items 21 - 40 were specified as subtest 2. The output file of BILOG 3.11 in this case contained two separate ability estimates, one for each subtest.

Multidimensional Item Response Scoring Model

Total Test Score MIRM (T)

To obtain ability estimates for the multidimensional compensatory scoring model, the Normal Ogive by Harmonic Analysis Robust Method (NOHARM III) (Fraser & McDonald, 2003) was used to estimate the multidimensional item parameters. The NOHARM program uses the non-linear factor analytic approach to estimate item parameters in exploratory and confirmatory modes. In this study the original item parameters used to generate the data (see Table 2) were used in NOHARM as the starting

parameters values to ensure that the item parameter estimates and the resulting ability estimates were as accurate as possible and reflected the original design of the data structure. The item parameters were estimated using the confirmatory mode since the dimensional structure of the test was known. Unfortunately, NOHARM does not estimate the multidimensional abilities of examinees. The multidimensional item parameter estimates from NOHARM were used in the input file for the 2D-EAP, a DOS-based program (Luecht, 1992) to obtain the ability estimates of the examinees on each of the two dimensions simultaneously using the *expected a posteriori* estimation method. The total test score was obtained by summing the subtest scores on each of the two dimensions. Obtaining the total test score this way resulted in the compensatory model; for example, high ability on one dimension can be compensated for low ability on the other dimension yielding a passing score.

Subtest scores MIRM (S)

The subtest scores on each of the two dimensions were obtained from the 2D-EAP program where separate ability estimate for each of the two dimensions were computed.

Given that the data were generated using the 2-parameter, 2-dimensional multidimensional compensatory model, the scores yielded by this model were considered to be the "true" examinees abilities. The classical test score model and the unidimensional item response models score estimates were compared to the ability values yielded by the multidimensional item response model.

Total Test Score Versus Subtest Scores

The scores yielded by the CTM, UIRM (T), and UIRM (S) scoring models were compared with the "true" scores yielded by the MIRM scoring model when total test score was considered. However, only scores yielded by the CTM and UIRM (S) were compared to scores yielded by the MIRM scoring model in the case of the subtest scores. The UIRM (T) was excluded when subtest scores were considered because only one ability estimate for the total test was produced.

Analyses

Since the ability estimates of examinees using the CTM, UIRM, and MIRM were expressed in different metrics, the scores were transformed to *T*-scores, with a mean of 50 and a standard deviation of 10 (Glass & Hopkins, 1996). This transformation of the scores allowed the comparison of the scores obtained from the different scoring models because the scores were in the same metric. In contrast, the mean of the ability distribution corresponding to the first dimension in Figure 8 was kept at 0.0 and the mean of the ability distribution corresponding to the second dimension was shifted to –1.0.

The agreement between the scores yielded by the MIRM scoring model and the scores yielded by each of the remaining three scoring models was assessed in four ways:

1. differences between the correlations between the examinees' scores yielded by MIRM and each of the other models for both total scores and subtest scores;
2. differences between the correlations between the examinees' subtest scores within scoring method;
3. root means square difference between the examinees' scores yielded by MIRM and each of the other models for both total scores and subtest scores; and

4. differences in rates of correct and incorrect classification of examinees based on scores yielded by MIRM and each of the other models for both total scores and subtest scores.

Correlation

Scatter plots were produced for randomly selected data sets in each condition to examine if the data meets the assumption of linearity. The correlation between the scoring methods was averaged across the 100 replications. For the total score, the scores yielded by the CTM, UIRM (T), and UIRM (S) were correlated with the scores yielded by the MIRM scoring model. Furthermore, for the above three scoring models, each subtest score corresponding to one of the two dimensions was correlated with the corresponding subtest score of the same dimension yielded by the MIRM scoring model. This generated two combinations for D1 and two combinations for D2. The correlation coefficient was obtained across the 100 replications. To obtain the mean correlation, the Pearson correlation coefficients were transformed to Fisher z:

$$z_{xy} = 0.5 \ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right),$$

where

z_{xy} is the Fisher z,

ln is the natural logarithm, and

r_{xy} is the Pearson correlation coefficient.

The mean of the transformed correlations was then computed. The mean is given by (Zimmerman *et al.*, 2003):

$$\bar{z} = \frac{\sum_{r=1}^{100} z_{xy}}{100}.$$

The mean was then transformed back to Pearson *r* by:

$$\bar{r} = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}.$$

Correlation Between Examinees' Subtest Score

The correlation between the subtest scores produced by the CTM, UIRM (S) scoring methods was compared to the correlations between the subtest scores produced by the MIRM scoring method to see the degree to which the input values were recovered. The Pearson correlation between the subtest scores was used. As previously discussed, the correlation values across the 100 replications were transformed to Fisher z. The mean of these correlations was computed and transformed back to Pearson's r.

Root Mean Square Difference

For each replication, the Root Mean Square Difference (RMSD) was used to examine if the scores yielded by the CTM and UIRM scoring models agreed with the scores yielded by the MIRM scoring model. The RMSD is given by:

$$RMSD_{ab} = \sqrt{\frac{\sum_{r=1}^{100}\sum_{j=1}^{2000}(T_a - T_b)^2}{(100)(1999)}},$$

where

T_a is the transformed score produced by scoring method *a*, and

T_b is the transformed score produced by scoring method *b*.

Classification

The classification of examinees was based on a cut-score of –1.0 standard deviations below the mean of the *T*-scores distribution. This cut-score translates into a score of 40 on the transformed score scale. For the total score, examinees who scored below 40 were classified as non-proficient and were placed in Group 1 whereas examinees who scored 40 or above were classified as proficient and were placed in Group 2.

For the subtest scores, examinees who had a combined score on the two dimensions below 80 were classified as non-proficient and were placed in Group 1. Examinees who had a combined score on the two dimensions of 80 or above were classified as proficient and were placed in Group 2.

The procedure of classifying examinees based on subtest scores was used because of the compensatory nature of the model. In other words, being high on one dimension can compensate for being low on another dimension. The categorization of examinees into Groups 1 and 2 using the subtest scores was based on Allan and Yen's (1979) rule 3 for classification using the compensatory model. Since this rule is the most frequently used method for prediction and admission decisions in education, it was adopted in this study. A pictorial representation of the compensatory model used in the present study is provided in Figure 9. Examinees whose scores fell in the areas a, e, and f were classified as Group 1 whereas examinees whose scores fell in the areas b, c, and d were classified as Group 2.

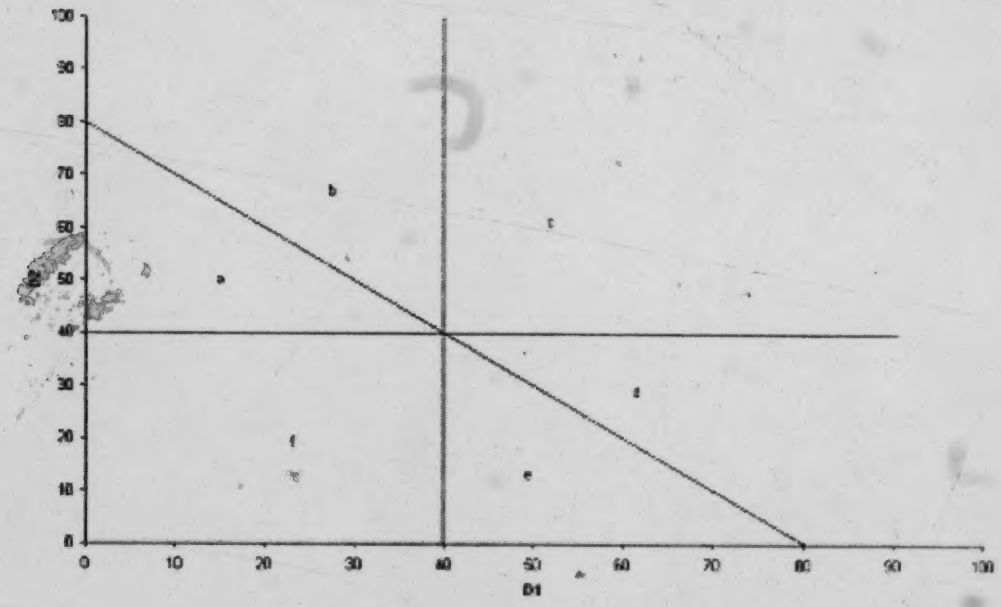


Figure 9. Cut – Score for Dimensions 1 and 2.

CHAPTER 4: RESULTS AND DISCUSSION

The purpose of the present study was to assess the degree to which the CTM, UIRM (T), and UIRM (S) scoring models were able to recover the scores obtained using the MIRM scoring model when the dimensionality of the test was known to be two. To address this purpose, the simulation study described in the previous chapter was conducted. The factors considered included:

1. correlation between examinees abilities on the two dimensions: 0.0, 0.3, 0.6, and 0.9;
2. mean differences on the two dimensions: equal (0,0) and unequal (0, -1);
3. factor complexity: simple and complex; and
4. type of score reported; total score and subtest score.

The simulation design corresponds to a 4 x 2 x 2 x 2 fully crossed design, yielding 32 conditions. The sample size of examinees for each simulation was set at 2,000 and the number of replications was set at 100.

The results of the assessment of agreement analyses are presented in this chapter. First, the results of a series of preliminary analyses conducted to ensure that the agreement results were accurate are presented. This is then followed by a presentation of the results for each of the four agreement analyses.

Preliminary Analyses

The first preliminary analysis involved checking to see that the data that entered each computer program was correct. The MULTISIM provided as part of its output a copy of the input file. For each condition, the copy of the input file was checked to ensure that the program read the input file correctly. No errors were found. The first two

examinees records are provided as part of the BILOG 3.11 output. Two data sets were randomly selected from each of the conditions defined by correlation, mean of ability distributions, and factor structure. The first two records in each were compared with the input data matrix. No differences were found. The second test involved comparing the CTM examinee scores obtained by SPSS and by BILOG 3.11. Again no differences were found. In the case of MIRM, the input parameters consisted of the initial values used to create the data sets. These values were compared to the read by NOHARM III. No differences were found. It was therefore concluded that the data files were read correctly.

The scores produced for each condition were then examined for meaningfulness. The scores were not produced for the set of conditions in which the correlation between examinee abilities was 0.9 and the factor structure was complex. The factors were ill-defined in these cases (T. Rogers, personal communication, July 8, 2005; R. P. McDonald, personal communication, July 13, 2005). Many values of the items *a*-parameters for these cases were negative and the estimated unique variance ranged from 0.36 to 0.64. Consequently there are no results for these conditions.

Third, prior to examining the correlational agreement between scores, the bi-variate distributions were examined for linearity. One data set was randomly selected from each of the remaining conditions. The best fitting line suggested that the relationship between the MIRM scores and each of the CTM, UIRM (T), and UIRM (S) scores was non-linear, particularly in the tails of the distribution. An example consisting of the MIRM scores and the CTM scores for the 9th replication of the condition in which the mean examinee ability on each dimension was the same, the correlation between examinee abilities was zero, there was simple structure, and a total score was computed

is provided in Figure 10. As shown, the best fitting line curves up in the left of the lower tail and down in the right of the upper tail. In between, the line was straight. This behavior was consistent across conditions and replications.

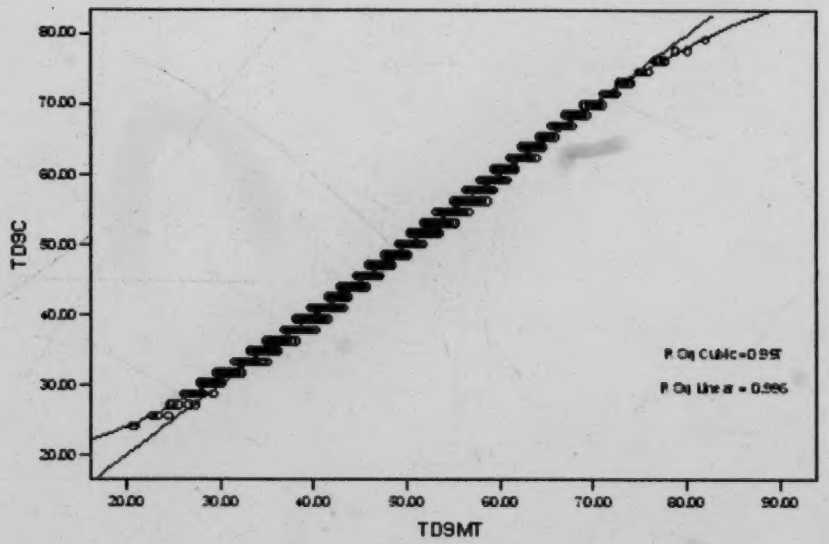


Figure 10. Scatter Plot of CTM and MIRM – 0000ST.

Consequently polynomials were fit to the data. The results revealed that the distribution was best fit by a cubic function. However, as described above, the departure from linearity was in the tails. Therefore, correlations between the examinee scores yielded by the MIRM scoring model and scores yielded by the other scoring models were computed for both the linear and cubic functions. The results are reported in Table 3 for the replication shown in Figure 10.

Table 3

Proportion of Explained Variance for Linear and Cubic Functions

		Statistics			
Function	R Square	F	df1	df2	Sig.
Linear	0.996	443610.39	1	1998	0.000
Cubic	0.997	203912.09	3	1996	0.000

The R square values were essentially identical (see Table 3). The same results were observed for other conditions. The departure from linearity in the tails had no appreciable effect on the correlation. Therefore, Pearson product moment correlations were used to assess correlational agreement.

Correlational Agreement

The product moment correlations were used to assess the degree to which the examinees were similarly ranked using the scores yielded by the MIRM scoring model and each of the CTM, UIRM (T), and UIRM (S) scoring models. The results using the total score are presented first. The results for the subtest or dimensions are then provided.

The following coding scheme is used to identify the conditions. Each code contains four numbers followed by two letters. The first two codes contain the dimension means: 00 in the case of equal means and 0-1 in the case of unequal means. The next two codes contain the input correlations between examinee abilities. For example, 00 indicates zero correlation while 03 indicates a correlation of 0.3. The first letter refers to the factor structure: "S" for simple and "C" for complex. The second letter refers to the type of score: "T" for total score and "S" for subtest scores. For example, the code for the first condition in Table 4 means equal means (00), input correlation of zero (00), simple structure (S), and total score (T).

Table 4 contains the correlations between total scores for the 14 conditions in which the total scores were used and that produced interpretable results. As shown in this table, all of the correlations were at least 0.99. The 2000 examinees were essentially ranked the same using the scores obtained by the MIRM scoring model and the CTM and each of the UIRM scoring models.

Table 4

Mean Correlations for Total Test Score

	Scoring Method		
Condition	CTM - MIRM	UIRM (T) - MIRM	UIRM (S) - MIRM
0000ST	0.998	0.998	0.999
0003ST	0.998	0.999	0.999
0006ST	0.998	0.999	0.999
0009ST	0.998	0.999	0.999
0-100ST	0.996	0.996	0.998
0-103ST	0.996	0.998	0.999
0-106ST	0.996	0.999	0.999
0-109ST	0.996	0.999	0.999
0000CT	0.996	1.000	0.999
0003CT	0.995	1.000	0.998
0006CT	0.995	0.999	0.998
0-100CT	0.991	1.000	0.998
0-103CT	0.991	0.999	0.998
0-106CT	0.991	0.999	0.997

Table 5 contains the correlations between the subtest scores for the 14 conditions in which the subtest scores were used and that produced interpretable results. As shown in this table, all of the correlations were at least 0.99 when the test structure was simple. However, the correlations were smaller when the test structure was complex, ranging from 0.906 (0006CS) to 0.959 (0-100CS). The lowest correlations values were reported for the conditions where the correlation between the two abilities was 0.6 and the test

structure was complex (e.g. 0.906, 0.910, 0.917, and 0.926). Nevertheless, the correlations reported for the complex structure were still very high. The 2000 examinees were essentially ranked the same using the subtest scores obtained by the MIRM scoring model and the CTM and the UIRM (S) scoring models.

Table 5

Mean Correlations for Subtest Scores

	Scoring Method			
Condition	CTM1 - MIRM1	CTM2 - MIRM2	UIRM(S)1 - MIRM1	UIRM(S)2 - MIRM2
0000SS	0.993	0.993	0.993	0.994
0003SS	0.993	0.993	0.994	0.994
0006SS	0.994	0.994	0.994	0.994
0009SS	0.995	0.995	0.995	0.995
0-100SS	0.992	0.990	0.993	0.995
0-103SS	0.993	0.990	0.994	0.995
0-106SS	0.994	0.990	0.994	0.996
0-109SS	0.994	0.991	0.995	0.996
0000CS	0.953	0.953	0.952	0.952
0003CS	0.947	0.945	0.946	0.945
0006CS	0.910	0.906	0.909	0.907
0-100CS	0.955	0.952	0.959	0.959
0-103CS	0.951	0.946	0.955	0.953
0-106CS	0.922	0.912	0.926	0.917

Correlations between Subtests within Scoring Method

The correlations between the subtest scores produced by the CTM and UIRM (S) scoring models were compared to the correlations between the subtest scores produced by the MIRM scoring model to see the degree to which the input values were recovered. The results of these analyses are presented in Table 6. As shown, the pattern of recovery is complex across and within scoring method.

Table 6

Mean Correlations between Subtests within Scoring Methods

	Scoring Method		
Condition	CTM	UIRM (S)	MIRM
0000SS	0.264	0.266	0.083
0003SS	0.462	0.466	0.305
0006SS	0.633	0.635	0.511
0009SS	0.783	0.782	0.701
0-100SS	0.237	0.235	0.075
0-103SS	0.437	0.438	0.295
0-106SS	0.611	0.612	0.499
0-109SS	0.764	0.764	0.690
0000CS	0.787	0.786	0.473
0003CS	0.837	0.832	0.572
0006CS	0.870	0.862	0.619
0-100CS	0.763	0.756	0.452
0-103CS	0.821	0.811	0.556
0-106CS	0.860	0.846	0.608

Generally, the MIRM scoring method reproduced the input correlations most closely. In the case of simple structure and equal dimension means, the fit between the mean MIRM correlations between the dimension scores and the corresponding input correlations was good for correlation values of 0.6 and below. When the dimension means were not equal, the fit was good for correlations less than 0.6. In the three cases where there was a difference, the input correlations were underestimated (0.701 (0009SS), 0.499 (0-106SS), and 0.690 (0-109SS)). In the case of complex structure, the pattern and values of the MIRM correlations were similar for both the equal and the unequal dimension mean conditions, with over estimation for the two lower correlations and near perfect fit at 0.6.

The patterns and values of the CTM and UIRM (S) correlations were similar to each other (see Table 6). In the case of simple structure, the best recovery of the input correlation was at 0.6. The input correlations below 0.6 were overestimated (e.g. 0.264 for 0000SS, CTM), while the input correlation of 0.9 was underestimated (e.g. 0.764 for 0-109SS). In the case of complex structure, the three input correlations were overestimated by an amount greater than that observed for simple structure (e.g. 0.264 for 0000SS, CTM versus 0.787 for 0000CS, CTM). As pointed out earlier in the discussion of the input correlation of 0.9 and complex structure (see page 69), these higher values are attributable to the overlap among test vectors that in turn led to difficulty in clearly defining the factors.

Root Mean Square Difference

As previously discussed in Chapter 3 (p. 65), the Root Mean Square Difference (RMSD) was computed across 100 replications to evaluate the absolute agreement between the scores yielded by the MIRM scoring method and each of the CTM, UIRM (T), and UIRM (S) scoring methods. Since the scores yielded by the four scoring methods were transformed to *T*-scores, the mean and the standard deviation of the scores are not presented because all the means and standard deviations across conditions were 50 and 10, respectively.

Total Score

The RMSD values of the total score for the simple and the complex structure are reported in Table 7 and graphically illustrated in Figure 11. As reported in Table 7 and illustrated in Figure 11, the mean differences between the scores for each of the three pairs CTM and MIRM, UIRM (T) and MIRM, and UIRM (S) and MIRM were below

one score point on the *T*- score scale (less than 10% of a standard deviation) for all conditions when the test structure was simple. The UIRM (S) and MIRM scoring methods had the smallest RMSD with the exception of two conditions 0003ST and 0006ST where the RMSD for the UIRM (T) and MIRM comparison was essentially the same. The RMSDs for the CTM and MIRM difference were the highest (e.g. 0.68 versus 0.51, 0000ST). The low values for the RMSD for the three pairs across all conditions with simple structure indicated that there was good agreement between the scores yielded by the MIRM scoring model and each of the CTM, UIRM (T), and UIRM (S) scoring models.

Table 7

Root Mean Square Difference for Total Test Score

	Scoring Method		
Condition	CTM - MIRM	UIRM (T) - MIRM	UIRM (S) - MIRM
0000ST	0.68	0.61	0.51
0003ST	0.65	0.40	0.43
0006ST	0.63	0.37	0.39
0009ST	0.62	0.40	0.36
0-100ST	0.91	0.92	0.59
0-103ST	0.89	0.60	0.49
0-106ST	0.87	0.54	0.42
0-109ST	0.85	0.53	0.38
0000CT	0.91	0.25	0.53
0003CT	0.96	0.28	0.57
0006CT	0.99	0.44	0.69
0-100CT	1.32	0.31	0.62
0-103CT	1.34	0.34	0.62
0-106CT	1.35	0.53	0.72

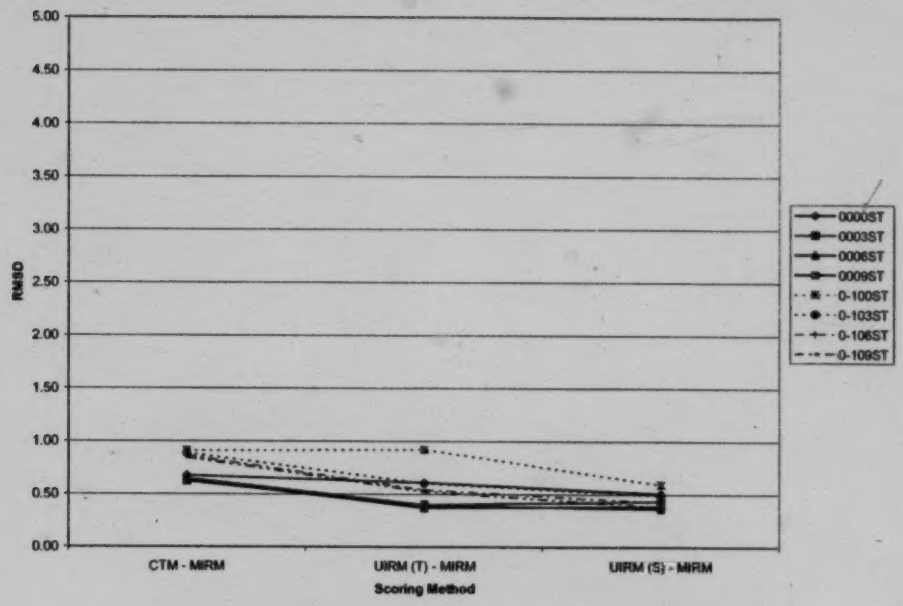


Figure 11. RMSD for Total Score – Simple Structure.

As reported in Table 7 and illustrated in Figure 12, the UIRM (T) scores appeared most closely to the MIRM scores when the test structure was complex. This finding is likely due to the high correlations among the dimension ability scores noted above. The next best fit was for the UIRM (S). The RMSD values were less than one score point for 15 of the 18 conditions. The remaining three RMSD values, which occurred for the comparisons of the CTM scores to the MIRM scores, unequal dimension means and complex structure, were less than 1.5 score points. In contrast to the simple structure conditions, the closest agreements occurred between the MIRM scores and the UIRM (T) scores for both equal and unequal dimension means.

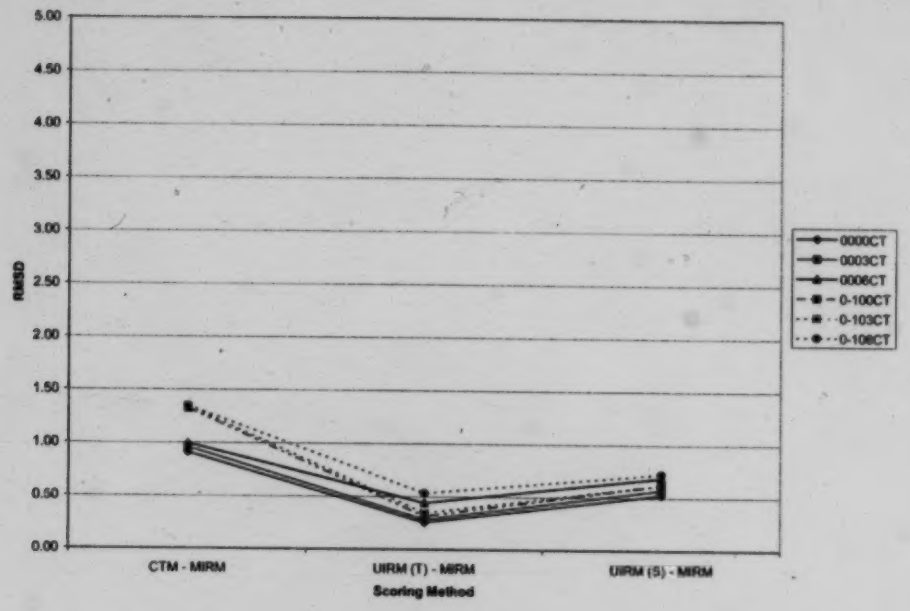


Figure 12. RMSD for Total Score – Complex Structure.

Subtest scores

The RMSD values of the subtest scores for the simple and the complex structure are reported in Table 8 and graphically illustrated in Figure 13. The first thing to note is that the values reported in this table are greater than the corresponding values reported in Table 7. Inspection of these findings revealed that there was closer agreement between the scores on both dimensions yielded by the UIRM (S) and the MIRM scoring methods than between the dimension scores yielded by the CTM and the MIRM scoring methods, particularly for the second dimension, unequal means when the test structure was simple. In the case of the first dimension, the RMSD values of the UIRM (S) – MIRM comparison varied between 0.99 (0009SS, 0-109SS) and 1.15 (0-100SS) while the RMSD values for the CTM – MIRM comparison varied between 1.02 (0009SS) and 1.24 (0-100SS). For the second dimension, the differences between the RMSD values were

greater. The RMSD values for the UIRM (S) – MIRM comparison varied between 0.88 (0-109SS) and 1.13 (0000SS) while for the CTM – MIRM comparison, the RMSD varied between 1.02 (0009SS) and 1.43 (0-103SS). The CTM scoring method was more sensitive to the shift in the mean of the ability distribution on the second dimension than the UIRM (S) scoring method.

Table 8

Root Mean Square Difference for Subtest Scores

	Scoring Method			
Condition	CTM1 - MIRM1	CTM2 - MIRM2	UIRM(S)1 - MIRM1	UIRM(S)2 - MIRM2
0000SS	1.22	1.22	1.14	1.13
0003SS	1.18	1.18	1.12	1.10
0006SS	1.11	1.11	1.07	1.05
0009SS	1.02	1.02	0.99	0.98
0-100SS	1.24	1.42	1.15	0.97
0-103SS	1.20	1.43	1.13	0.96
0-106SS	1.14	1.41	1.08	0.93
0-109SS	1.05	1.37	0.99	0.88
0000CS	3.07	3.08	3.10	3.10
0003CS	3.30	3.36	3.34	3.36
0006CS	4.30	4.39	4.34	4.38
0-100CS	3.01	3.11	2.88	2.87
0-103CS	3.15	3.30	3.02	3.08
0-106CS	4.03	4.28	3.95	4.18

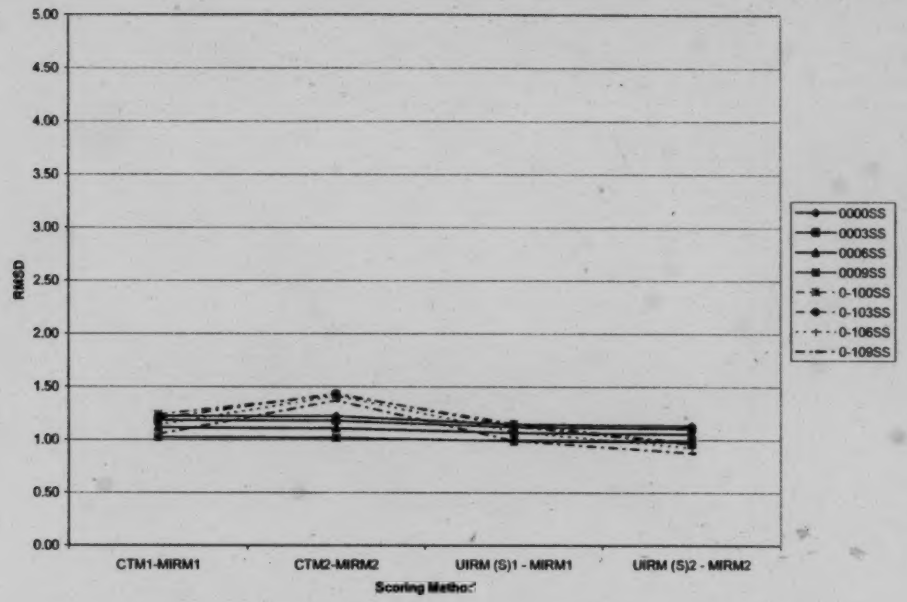


Figure 13. RMSD for Sub-Test Scores – Simple Structure.

For complex structure, the RMSD values are reported in Table 8 and graphically illustrated in Figure 14. As reported in Table 8, the RMSD values for complex structure were much higher than the RMSD values for simple structure. The lowest RMSD value (2.87, 0-100CS) for complex structure was considerably larger than the highest RMSD (1.43, 0-103SS) for simple structure. The RMSD values were less than 3.4 score points (less than 34% of a standard deviation) for 16 of the 24 values. For both the CTM and UIRM (S) the RMSD values were lower than 4.4 score points (less than 44% of a standard deviation) for the condition where the correlation between the two abilities was 0.6 for both equal and unequal means. Comparison of the RMSD when the subtests were recognized and when the subtests were ignored reveals that there were closer agreements when using total scores than when using subset scores. This is due to the more variations

in scores yielded by the MIRM scoring method when subtest scores were used. These variations were even more pronounced when the subtest scores were used and the test structure is complex. The increased variations were the result of the two factors becoming increasingly ill-defined as the correlation between abilities increased.

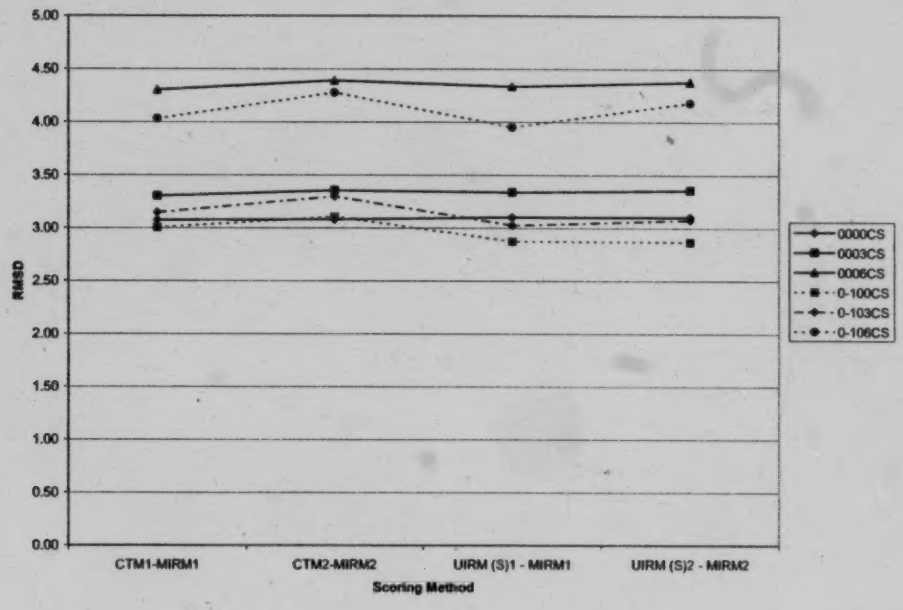


Figure 14. RMSD for Sub-Test Scores – Complex Structure.

Classification

Many testing programs establish performance standards and the corresponding cut-scores in the distribution of scores. For example, students can be classified as masters or non-masters. In Alberta, students are classified into one of three categories based on their performance on the provincial achievement tests at Grades 3, 6, and 9: met the standard of excellence, met the standard of acceptability, or did not meet the standard of acceptability. Consequently the ability of the CTM, UIRM (T), and UIRM (S) scoring

methods to recover the classifications based on scores yielded by the MIRM scoring method was assessed using the total score. The classification agreement of the CTM and UIRM (S) was also assessed using the two dimension scores.

The results of the classifications are reported in Tables 9 and 10 for the total score, and in Tables 11 and 12 for the subtest scores. Each table contains a series of two-way contingency tables. The rows correspond to the classifications based on the scores yielded by the MIRM scoring model. As mentioned earlier, the fit between the each of the CTM and UIRM scoring models and the MIRM scoring model for each condition was completed using 100 common samples for each condition. The fact that the common samples differed by condition accounts for the slight variations among the rows in the tables. The numbers in the last column represent the total number of Group 1 and Group 2 examinees identified using the MIRM scores. The columns correspond to the classifications based on the scores yielded by the CTM and the appropriate UIRM model. The principal diagonal of each contingency table contains the proportions of agreement for Group 1 and Group 2. The off-diagonal elements represent misclassifications. Lastly, the overall agreement is provided under the total heading for each table. For example, 303 and 1,697 examinees were classified into Group 1 and Group 2, respectively, using the MIRM model for condition 0000ST (See Table 9). Of the 303 examinees in Group 1, 98.4% were so classified by the CTM scoring model. Likewise, 98.4% of the Group 2 examinees were so classified by the CTM model. Thus, the overall classification agreement between the MIRM scoring model and the CTM scoring model was 98.4%.

As shown in Tables 9 and 10, the CTM, UIRM (T), and UIRM (S) scoring methods recovered well the classifications based on scores yielded by the MIRM scoring method when the total score was used regardless of the complexity of factor structure. The overall mean percentage of like decisions ranged between 95.7% and 99.0%. While the rates of misclassifications were low, there was a tendency for a greater percentage of Group 1 examinees to be misclassified. However, these percentages were no greater than 7%. Further, this finding need to be mediated by the smaller number of examinees classified in Group 1 than in Group 2. A shift of one student in Group 1 represents a greater shift in the percentages in row 1 of each contingency table than in the percentages in row 2.

Table 9

Classification of MIRM and CTM, UIRM (T), and UIRM (S) for Total Score - Simple Structure.

			Scoring Method									
			CTM		Total	UIRM (T)		Total	UIRM (S)		Total	N
Condition			1	2		1	2		1	2		
0000ST	MIRM	1	0.984	0.016		0.958	0.042		0.956	0.044		303
		2	0.016	0.984		0.006	0.994		0.008	0.992		1697
					0.984			0.976			0.974	
0003ST	MIRM	1	0.928	0.072		0.936	0.064		0.950	0.050		305
		2	0.012	0.988		0.008	0.992		0.006	0.994		1695
					0.958			0.964			0.972	
0006ST	MIRM	1	0.977	0.023		0.960	0.040		0.969	0.031		309
		2	0.012	0.988		0.002	0.998		0.005	0.995		1691
					0.983			0.979			0.982	
0009ST	MIRM	1	0.995	0.015		0.959	0.041		0.970	0.030		312
		2	0.005	0.985		0.001	0.999		0.003	0.997		1688
					0.990			0.979			0.983	
0-100ST	MIRM	1	0.936	0.064		0.925	0.075		0.967	0.033		299
		2	0.007	0.993		0.012	0.988		0.009	0.991		1701
					0.964			0.957			0.979	
0-103ST	MIRM	1	0.928	0.072		0.936	0.064		0.950	0.050		305
		2	0.012	0.988		0.008	0.992		0.006	0.994		1695
					0.958			0.964			0.972	
0-106ST	MIRM	1	0.970	0.030		0.961	0.039		0.973	0.027		309
		2	0.012	0.988		0.007	0.993		0.007	0.993		1691
					0.979			0.977			0.983	
0-109ST	MIRM	1	0.944	0.056		0.939	0.061		0.959	0.041		314
		2	0.008	0.992		0.002	0.998		0.004	0.996		1686
					0.968			0.968			0.978	

Table 10

Classification of MIRM and CTM, UIRM (T), and UIRM (S) for Total Score – Complex Structure.

			Scoring Method									
			CTM		Total	UIRM (T)		Total	UIRM (S)		Total	N
Condition			1	2		1	2		1	2		
0000CT	MIRM	1	0.994	0.006		0.979	0.021		0.978	0.022		304
		2	0.020	0.980		0.004	0.996		0.008	0.992		1696
					0.987			0.988			0.985	
0003CT	MIRM	1	0.971	0.029		0.974	0.026		0.979	0.021		308
		2	0.008	0.992		0.007	0.993		0.011	0.989		1692
					0.982			0.983			0.984	
0006CT	MIRM	1	0.943	0.057		0.940	0.060		0.941	0.059		316
		2	0.014	0.986		0.006	0.994		0.011	0.989		1684
					0.965			0.967			0.965	
0-100CT	MIRM	1	0.968	0.032		0.987	0.008		0.967	0.033		303
		2	0.014	0.986		0.013	0.992		0.011	0.989		1697
					0.977			0.990			0.978	
0-103CT	MIRM	1	0.971	0.029		0.974	0.026		0.979	0.021		308
		2	0.008	0.992		0.007	0.993		0.011	0.989		1692
					0.982			0.983			0.984	
0-106CT	MIRM	1	0.943	0.057		0.940	0.060		0.941	0.059		316
		2	0.014	0.986		0.006	0.994		0.011	0.989		1684
					0.965			0.967			0.965	

The agreement results for the classification of examinees based on subtest scores and using the compensatory model described above were similar to the agreement results obtained using the total score. The percentages of overall mean hit rate ranged from 97.0% to 99.0% across all conditions for the two subtest scoring models. As for the total scores, while the rate of misclassifications was low, there was a tendency for a greater percentage of Group 1 examinees to be misclassified. Again this may be attributed to the smaller number of examinees who were classified as Group 1.

Table 11

Classification of MIRM, CTM, and UIRM (S) for Subtest Scores – Simple Structure.

			Scoring Method						
			CTM		Total	UIRM(S)		Total	N
Condition			1	2		1	2		
0000SS	MIRM	1	0.998	0.002		0.997	0.003		165
		2	0.018	0.982		0.018	0.982		1835
					0.990			0.989	
0003SS	MIRM	1	0.997	0.003		0.993	0.007		212
		2	0.021	0.979		0.015	0.985		1788
					0.988			0.989	
0006SS	MIRM	1	0.994	0.006		0.988	0.012		251
		2	0.022	0.978		0.011	0.989		1749
					0.986			0.988	
0009SS	MIRM	1	0.992	0.008		0.982	0.018		284
		2	0.019	0.981		0.007	0.993		1716
					0.986			0.987	
0-100SS	MIRM	1	0.963	0.037		0.984	0.016		159
		2	0.015	0.985		0.015	0.985		1841
					0.974			0.984	
0-103SS	MIRM	1	0.950	0.050		0.984	0.016		204
		2	0.010	0.990		0.013	0.987		1796
					0.970			0.986	
0-106SS	MIRM	1	0.954	0.046		0.985	0.015		245
		2	0.009	0.991		0.011	0.989		1755
					0.973			0.987	
0-109SS	MIRM	1	0.967	0.033		0.984	0.016		282
		2	0.009	0.991		0.009	0.991		1718
					0.979			0.987	

Table 12

Classification of MIRM, CTM, and UIRM (S) for Subtest Scores – Complex Structure.

			Scoring Method						
			CTM		Total	UIRM (S)		Total	N
Condition			1	2		1	2		
0000CS	MIRM	1	1.000	0.000		1.000	0.000		240
		2	0.040	0.960		0.032	0.968		1760
					0.980			0.984	
0003CS	MIRM	1	1.000	0.000		0.999	0.001		261
		2	0.038	0.962		0.025	0.975		1739
					0.981			0.987	
0006CS	MIRM	1	0.999	0.001		0.991	0.009		276
		2	0.040	0.960		0.021	0.979		1724
					0.980			0.985	
0-100CS	MIRM	1	0.996	0.004		1.000	0.000		231
		2	0.021	0.979		0.030	0.970		1769
					0.987			0.985	
0-103CS	MIRM	1	0.989	0.011		0.998	0.002		254
		2	0.027	0.973		0.030	0.970		1746
					0.981			0.984	
0-106CS	MIRM	1	0.960	0.040		0.982	0.018		268
		2	0.016	0.984		0.025	0.975		1732
					0.972			0.979	

CHAPTER 5: SUMMARY AND CONCLUSIONS

A summary of the study and the methods used is presented at the beginning of this chapter. The key findings are then summarized followed by the identification of the limitations of the study. Conclusions are presented next followed by the implications and recommendations for future research.

Summary of the Study

Since test scores are used as evidence to make important decisions about examinees, the scores must be determined accurately in order for sound decisions to be made. The purpose of the present study was to assess the degree to which the classical test score model (CTM), the 2-parameter unidmensional item response model with total calibration (UIRM (T)), and with separate calibration (UIRM (S)) were able to recover the scores obtained using the MIRM scoring model when the dimensionality of the test was known to be two. Since the MIRM scoring model is the appropriate model to score multidimensional data, the scores yielded by this model were considered to be "true" scores. To address the purpose of this study, a simulation study was conducted. The factors considered included:

1. correlation between examinees abilities on the two dimensions: 0.0, 0.3, 0.6, and 0.9;
2. differences on the two dimensions: equal (0,0) and unequal (0, -1);
3. factor complexity: simple and complex; and
4. type of score reported: total score and subtest score.

The simulation design corresponded to a 4 x 2 x 2 x 2 fully crossed design, yielding 32 conditions. The sample size of examinees for each simulation was set at 2,000 and the number of replications was set at 100.

The agreement between the scores yielded by the MIRM scoring model and the scores yielded by each of the remaining three scoring models was assessed in four ways:

1. differences between correlations between the examinees' scores yielded by MIRM and each of the other models for both total scores and subtest scores;
2. differences between the correlations between the examinees' subtest scores within scoring method;
3. root means square difference between the examinees' scores yielded by MIRM and each of the other models for both total scores and subtest scores; and
4. differences in rates of correct and incorrect classification of examinees based on scores yielded by MIRM and each of the other models for both total scores and subtest scores.

Summary of Key Findings

At the outset, it was found that many of the multidimensional *a* parameter estimates were negative when the correlation between the two abilities was 0.9 and when the test structure was complex. This finding was due to the ill-defined factors. As a result, this condition was eliminated from the analysis.

The following findings are organized in terms of the above four assessments.

1. For the total test score, the MIRM and each of the CTM, UIRM (T), and UIRM (S) ranked examinees similarly for both simple and complex structure. The correlations between the scores were at least 0.99 across all conditions.

For the sub-test scores, the MIRM, CTM, and UIRM (S) also ranked examinees similarly when the test structure was simple. The correlation between the sub-test scores was also at least 0.99. However, the correlation between the subtest scores was lower for the complex structure and ranged from 0.91 to 0.96. Although the correlations were lower when the test structure was complex, these correlations are still quite high.

2. The pattern of recovery of the correlations between the subtest scores was complex across and within the scoring methods. The MIRM reproduced the correlations between the subtest scores most closely for simple structure and equal dimension means when the correlation was 0.6 and below. However, the MIRM reproduced the correlations most closely for simple structure and unequal dimension means when the correlation was 0.3 and below.

   For complex structure, the MIRM reproduced the correlations at 0.6 for both equal and unequal mean conditions. The patterns and values of the CTM and the UIRM (S) correlations were similar to each other where these methods recovered correlations of 0.6 when the test structure was simple. In the case of complex structure, the three input correlations were overestimated by an amount greater than that observed for simple structure. The results in this case were attributable to the overlap among test vectors that in turn led to difficulty in clearly defining the factors.

3. For all conditions when the test structure was simple, the RMSD values between the MIRM scores and each of the CTM, UIRM (T), and UIRM (S) scores were less than one score point on the *T*-score scale when the total test score was used

indicating high agreement between the scoring methods. The strongest agreement, however, was between the UIRM (S) and the MIRM scoring models.

For the complex structure, the strongest agreement was between the UIRM (T) and the MIRM scoring methods. However, the differences in RMSD values between the UIRM (S) and the UIRM (T) were small indicating that there was also a strong agreement between the UIRM (S) and the MIRM.

For the subtest scores and simple structure, all of the RMSD values for the UIRM (S) and MIRM comparison were 1.15 scores or less indicating a good agreement between the two scoring methods. For the CTM – MIRM comparison, the RMSD values were less than 1.25 scores. The CTM was more sensitive to the shift in the mean of the ability distribution on the second dimension where the RMSD values increased to less than 1.5 points.

For the complex structure, the agreement between the MIRM and each of the CTM and UIRM (S) was low where the RMSD values were at least 3 score points. These results were due to the previously mentioned ill-defined factors.

4. The classification results revealed high rates of agreement based on the mean percentage between the MIRM and each of the CTM, UIRM (T), and UIRM (S) methods. The mean percentage of agreement ranged from 0.96 to 0.99 across all conditions for both simple and complex structure when the total score or the subtest scores were used. The percentage of examinees that were misclassified was small across all conditions. Therefore, all the scoring methods classified examinees similarly.

Limitations of the Study

This study was limited to using the two-parameter unidimensional model and the two-parameter compensatory multidimensional model. Therefore, before the results can be generalized to other unidmensional and multidimensional models, further research is needed. The difference between the values of the item parameters was low for both simple and complex structure. The low item parameter values led to the item vectors being similar in measuring the two dimensions especially when the test structure was complex. As a result, the communalities were very small which led to factor under identification. Hakstian, Rogers, and Cattell (1982) concluded that when the communalities were moderate to low, identifying the number of factors became more problematic. The low communalities led to ill-defined factors, thus procedures for identifying the factors became less reliable. Another limitation is that the number of dimensions was restricted to two and the sample size was restricted to 2000 examinees. Further, since only dichotomously- scored items were used, the results do not necessarily generalize to polytomously-scored items. The mean of the ability distribution on the second dimension was shifted to −1.0 standard deviations, therefore, the results do not generalize to more extreme values. Finally, only one cut-score was used to separate the examinees into two different groups. Hence, the classification results are based only on this cut-score.

Conclusions

The comparison between the MIRM scoring method and each of the CTM, UIRM (T), and UIRM (S) scoring methods revealed that the results were equivocal. For simple structure, all the methods ranked examinees similarly across all conditions when the total

score or the subtest scores were used. The correlations were at least 0.99. Likewise, the RMSD results suggest that there was good agreement between the scores yielded by the MIRM and each of the CTM, UIRM (T), and UIRM (S). The RMSD values were lower than 1.50 points (less than 15% of a standard deviation) on the *T*-score scale. Lastly, the four scoring methods also classified examinees similarly where the mean percentage of classifications ranged between 96% and 98%.

For complex structure, the MIRM and each of the CTM, UIRM (T), and UIRM (S) ranked examinees similarly across all conditions when the total score was used. The correlations were at least 0.99. However, the correlations between scores yielded by the MIRM and each of the CTM and UIRM (S) were lower for the subtest scores. The subtest correlations ranged between 0.91 and 0.95 indicating that the ranking of examinees on the subtests was less similar across the scoring methods than when the total score was used. RMSD values were less than 1.50 score points when the total test score was used indicating good agreement between the scores yielded by MIRM and each of the CTM, UIRM (T), and UIRM (S). In contrast, while the RMSD values were less than half a standard deviation, they were approximately three times larger when the subtest scores were used, suggesting again that there was less score agreement between the MIRM and the other scoring methods at the subtest level. However, this apparent lack of agreement did not adversely influence the classification of students. The classification results revealed that all the methods using both the total scores and the subtests scores classified examinees similarly. The mean percentages agreement ranged from 96% to 98%. The high rates of classification agreement contradict the high RMSD

values for subtest scores. It appears that the apparent lack of agreement was not sufficiently large, particularly around the cut-score considered in the present study (-1.0).

Taken together, the results of this study suggest that the use of multidimensional, unidimensional, and number-right scoring will not lead to differences when total test scores derived from multidimensional tests with simple structure are reported. The results are less clear with subtest scoring when the structure is complex, but this may be an artifact of low communality that led to factor confusion in the present study (see limitations). Given that the number-right scoring method is relatively easy to apply, does not require large sample sizes, and is more familiar and easier to explain to teachers and the general public than the item response based methods, the findings of the present study support the use of number-right scoring when the factor structure is relatively simple, which is the case for most tests in current use.

Implications for Future Practice

Based on the results of this study, there appears to be no need for large-scale testing programs (e.g. state, provincial, national, international) to change their scoring procedures. This recommendation is contingent on the use of multidimensional tests whose structure is simple. Large-scale testing programs that use the classical test score model to obtain their scores should continue to do so. Likewise, testing agencies that use either a unidimensional or multidimensional item response scoring model need not change their scoring method. The restriction of this recommendation is based on the findings in this dissertation that the results were equivocal for multidimensional tests with complex structure. The lack of a clear picture prevented recommending one scoring

method over another in the case of multidimensional tests with complex structure. However, as pointed out in the conclusion, well constructed tests tend to exhibit good simple structure.

Implications for Future Research

Future research is needed to address the problem of low communality. This can be done by increasing the discrimination values of the items on the dimension to which each item belongs. At the same time, the difference between the item parameters measuring the two dimensions should be increased. Higher discriminating parameters and a larger difference between the discriminating parameters on each dimension would create well defined factors.

Further, in this study the discrimination parameters for items measuring dimension 1 and for items measuring dimension 2 were kept constant (0.60) to avoid obtaining scores that were higher on one dimension because that dimensions had higher discriminating item parameters than the second dimension. More variable discrimination parameters with the same mean for each dimension should be investigated to assess the generalizability of the findings of the present research. This will allow the simulation to be more realistic since most tests have variable discrimination parameters among the high values along each dimension in a multidimensional case.

Although this study compared the scoring methods under different conditions, the scoring methods should also be compared using different mean values for the distributions of abilities. For example, more extreme values of the mean ability distribution on one of the dimensions as opposed to another. Clearer differences between the scoring methods may emerge because the classical scoring model showed

some sensitivity toward shifts in the mean ability distribution. Further, the shape of the ability distributions were similar, the effect of skewed distribution on scoring may be explored.

In this study, the test was assumed to be two-dimensional, in reality tests may have more than two dimensions. Therefore, the effects of using higher numbers of dimensions on scoring could be explored.

The sample size of 2000 examinees was used in this study. Large-scale testing agencies often deal with large sample of students. However, these testing agencies also study subgroups of examinees where the sample size is smaller than 2000. The effects of using sample sizes of 1000 or 1500 examinees on scoring could be entertained.

Further research is needed at different cut-scores. Only one cut-score was considered in the present study. This situation corresponds to a master/non-master or go/no-go decision. However many testing programs now use more than one cut-score to categorize students into more than two classes.

Finally, since simulation was used in this study, it would also be essential to apply the findings to real test data to find out if these results translate to real situations. In addition, extending this study to include polytomously-scored items would be important since many testing programs are increasingly using this type of test items.

References

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, 7, 255 – 278.

Ackerman, T. A. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, 20, 311 – 330.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, 22, 37 - 53.

Anderson, J. O. (1999). Does complex analysis (IRT) pay any dividends in achievement testing? *The Alberta Journal of Educational Research*, XLV, 344 – 352.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6, 431 – 444.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). *TESTFACT 4.0* [Computer Software]. Chicago, IL: Scientific Software International.

Douglas, J., Kim, H. R., Roussos, L., Stout, W., & Zhang, J. (1999). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992, administrations.* (Statistical Report No. 95 – 05). Newton, PA: Law School Admission Council.

Embretson, S.E. (1985). Multicomponent models for test design. In S.E. Embretson (Ed.) *Test design: Developments in psychology and psychometrics*. New York: Academic Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawerence Erlbaum & Associates, Inc.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item / person statistics. *Educational and Psychological Measurement*, 58, 357 – 381.

Fraser, C., & McDonald, R. P. (2003). *NOHARM 3.0: A windows program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England.

Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology*. Needham, MN: Allyn & Bacon.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 18*, 225 – 239.

Hakstian, R. A., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number of factors rules with simulated data. *Multivariate Behavioral Research, 17*, 193 – 219.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. NewBury Park, CA: Sage.

Kelly, T. L. (1942). The reliability coefficient. *Psychometrika, 7*, 75 – 83.

Lord, F. M. (1952). A theory of test scores. *Psychometric Monographs*, Whole No. 7.

Lord, F. M. (1980). *Application of item response theory to practical test problems*. Hillsdale, NJ: Lawrence Erlbaum & Associates, Inc.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, Mass.: Addison – Wesley.

Law School Admission Test (1999). Newton, PA: Law School Admission Council.

Luecht, R. M. (2003). *Applications of multidimensional diagnostic scoring for certification and licensure tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (1992). *2D–EAP.* Bayesian estimation software program for computing two-dimensional expected a posteriori IRT proficiency scores, based on the multidimensional three-parameter normal ogive model. Iowa City, IA: ACT.

Luecht, R. M., & Miller, T. R. (1992). Unidimensional calibrations and interpretations of composite traits for multidimensional tests. *Applied Psychological Measurement*, 16, 279 – 293.

McKinley, R. L., & Reckase (1982). *The use of the general Rasch model with multidimensional item response data* (Research Report ONR 82 – 1). Iowa City, IA: American College Testing.

Mislevy, R. J. & Bock, R. D. (1990). *BILOG 3.0 user's guide.* Mooresville, IN: Scientific Software International.

Mislevy, R. J., & Bock, R. D. (1997). *BILOG 3.11: Item analysis and test scoring with binary logistic models.* Mooresville, IN: Scientific Software International.

Ndalichako, J., & Rogers, W. T. (1997). Comparison of finite state score theory, classical test score theory, and item response theory in scoring multiple-choice items. *Educational and Psychological Measurement*, 57, 580 – 589.

Principles for Fair Student Assessment Practices for Education in Canada. (1993). Edmonton, AB: Joint Advisory Committee.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Denmark Paedagogiske Institut. (Republished in 1980 by the University of Chicago Press of Chicago).

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement*, 9, 401 – 412.

Reckase, M. D. (1990). Unidimensional data from multidimensional tests and multidimensional data from unidimensional tests. Paper presented at the annual meeting of the American Educational Research Association.

Reckase, M. D., Ackerman T. A., & Carlson, J. E. (1988). Building a unidimensional test using multidimensional items. *Journal of Educational Measurement*, 25, 193 – 203.

Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied Psychological Measurement*, 14, 361 – 373.

Rogers, W. T., & Ndalichako, J. (2000). Number right, item response and finite state scoring: Robustness with respect to lack of equally classifiable options and item option independence. *Educational and Psychological Measurement,* 60, 5 – 19.

Rogers, W. T. (2000). *Test Theory*. Edmonton: University of Alberta. [course notes.]

Rulon, P. J. (1939). A simplified procedure for determining reliability. *Harvard Educational Review,* 9, 99 – 103.

Spearman, C. (1904). The Proof and measurement of association between two things. *American Journal of Psychology*, 15, 72 – 101.

Standards for Educational and Psychological Testing. (1999). Washington, DC: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293 – 325.

Stout, W. F. (1990). *MULTISIM* [Computer Software]. St. Paul, MN: Assessment Systems Corporation.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331 – 354.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum & Associates, Inc

Tate, R. L. (2002). Test dimensionality. In J. Tindal & T. M. Haladyna (Eds.), *Large scale assessment programs for all students: Development, implementation, and analysis*. Mahwah, NJ: Lawrence Erlbaum & Associates, Inc.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, 17, 89 – 112.

Thissen, D., & Wainer, H. (2001). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum & Associates, Inc.

Tomkowicz, J. T., & Rogers, W. T. (in press). The use of one-, two-, and three-parameter and nominal item response scoring in place of number-right scoring in the presence of testwiseness. *The Alberta Journal of Educational Research.*

Wainer, H., & Thissen, D. (1987). Estimating ability with the wrong model. *Journal of Educational Statistics*, 12, 239 – 368.

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement, 40*, 255 - 275.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). *CONQUEST* [Computer Software]. Melbourne, Australia: Australian Council for Educational Research.

Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicologica*, 24, 133 – 158.

Appendix A

Total Score

This appendix provides the scatter plots of the total score yielded by the MIRM and each of the CTM, UIRM (T), and UIRM (S). For each condition, a data set was randomly selected from the 100 replications. Figure 1 represents the scatter plot of the CTM – MIRM comparison for condition 0000ST where the scores from the 9th replication were used.

Figure 1: Scatter plot for CTM and MIRM - 0000ST

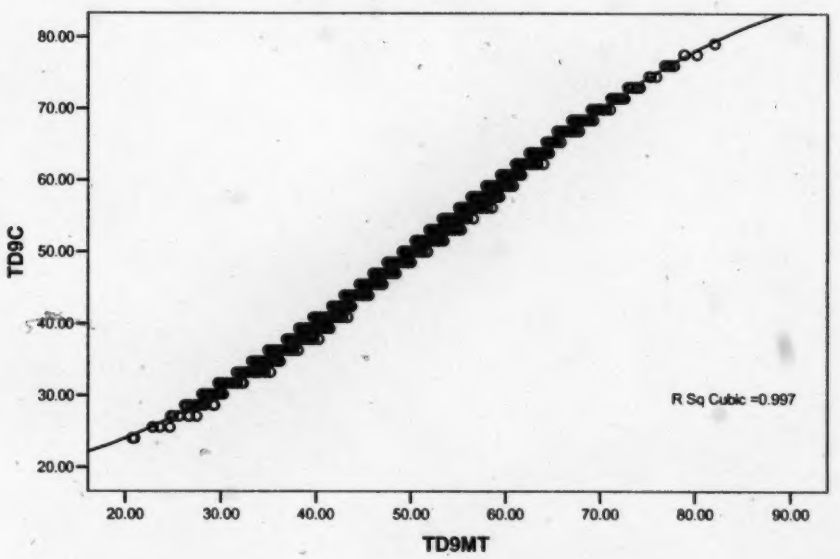


Figure 2: Scatter plot for UIRM (T) and MIRM - 0000ST

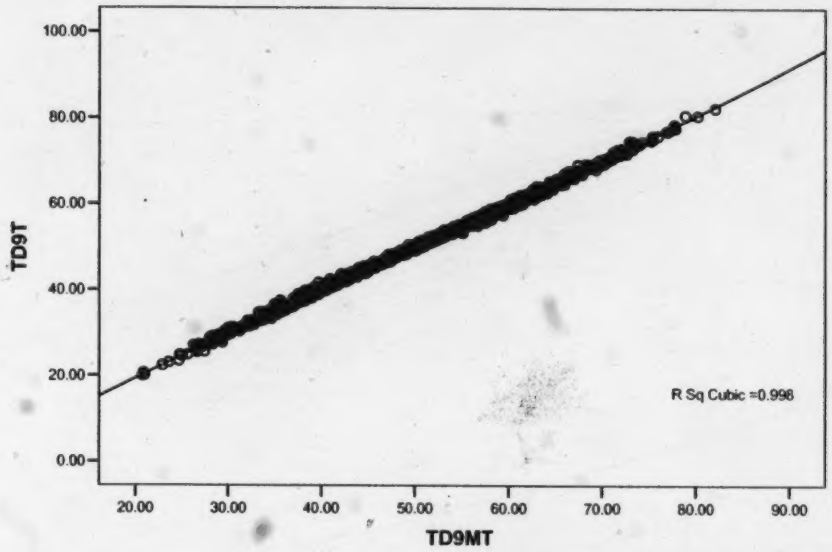


Figure 3: Scatter plot for UIRM (S) and MIRM - 0000ST

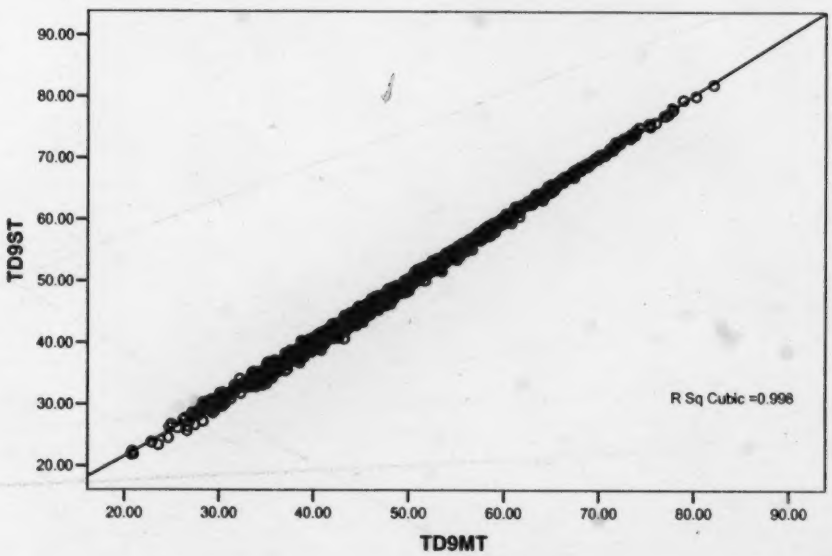


Figure 4: Scatter plot of CTM and MIRM - 0003ST

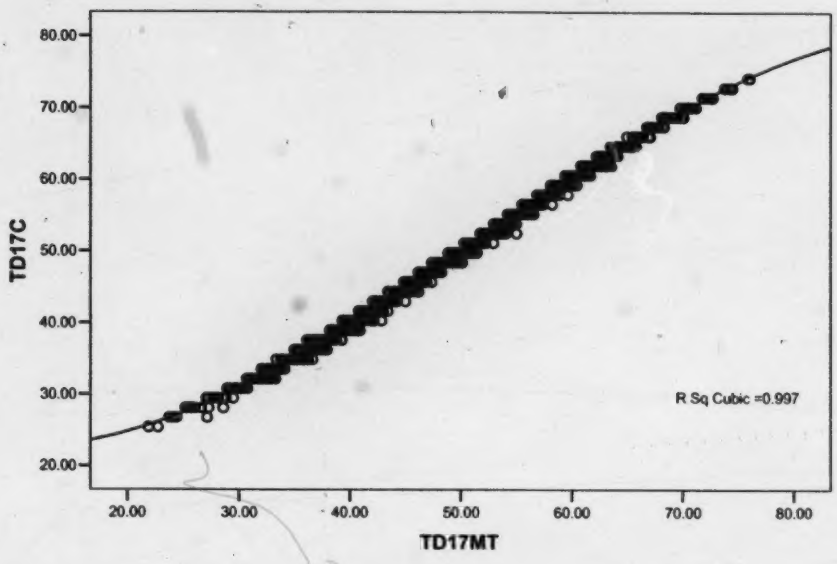


Figure 5: Scatter plot of UIRM (T) and MIRM - 0003ST

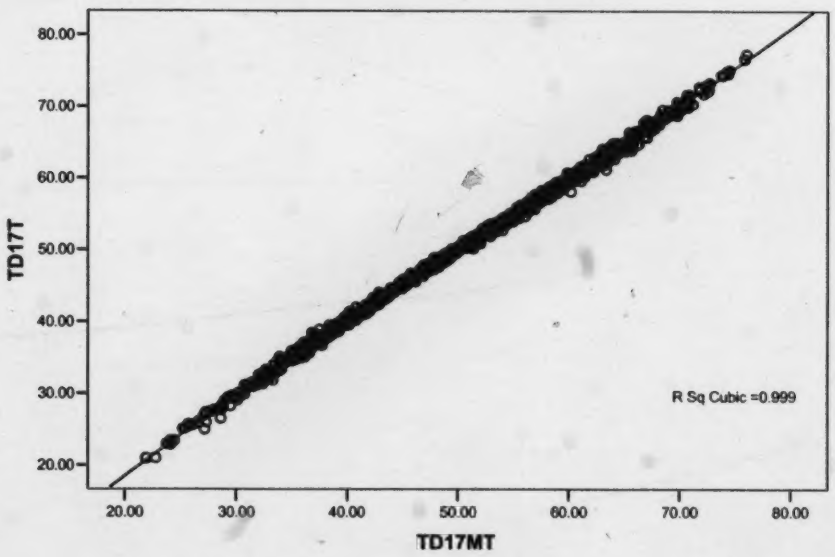


Figure 6: Scatter plot of UIRM (S) and MIRM - 0003ST

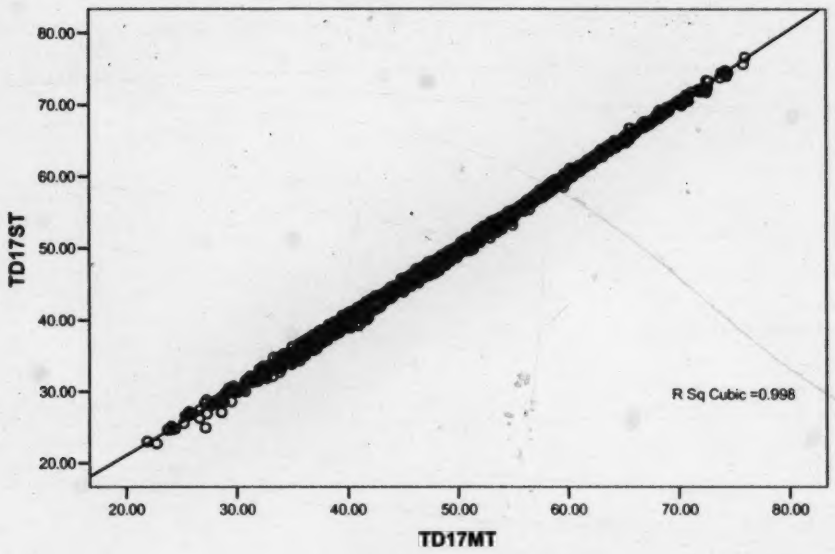


Figure 7: Scatter plot of CTM and MIRM - 0006ST

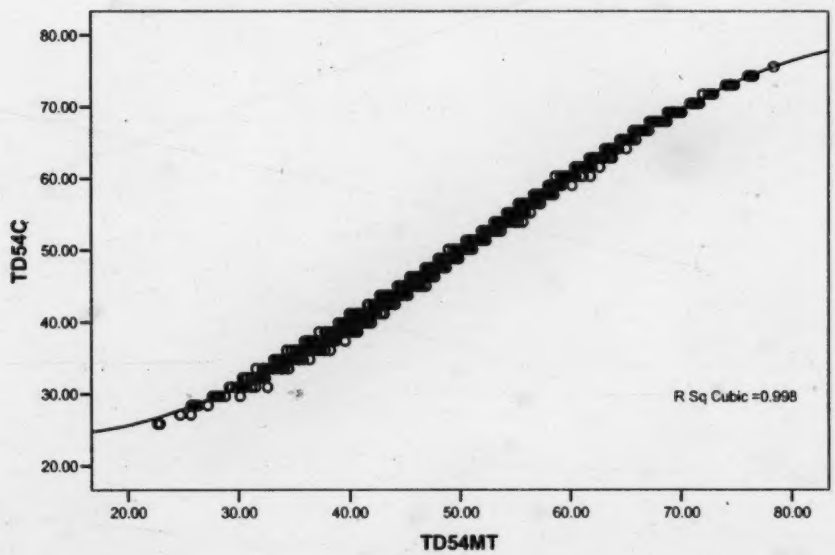


Figure 8: Scatter plot of UIRM (T) and MIRM - 0006ST

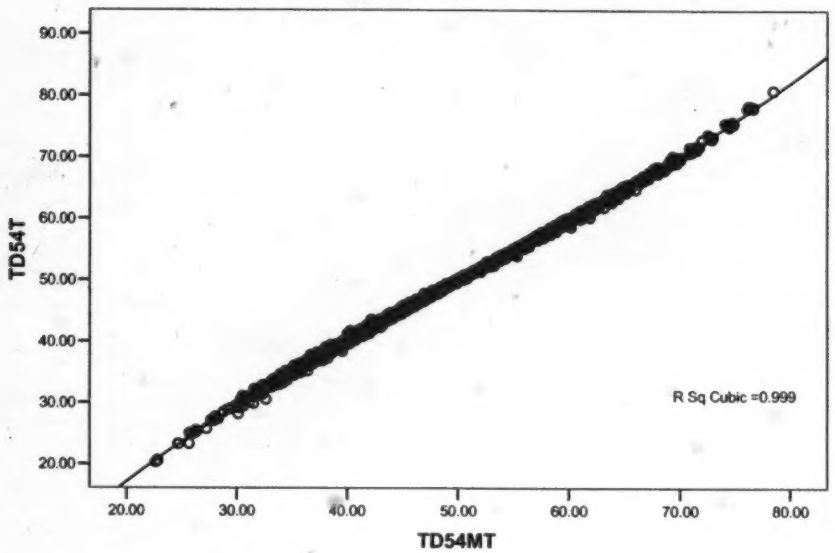


Figure 9: Scatter plot of UIRM (S) and MIRM - 0006ST

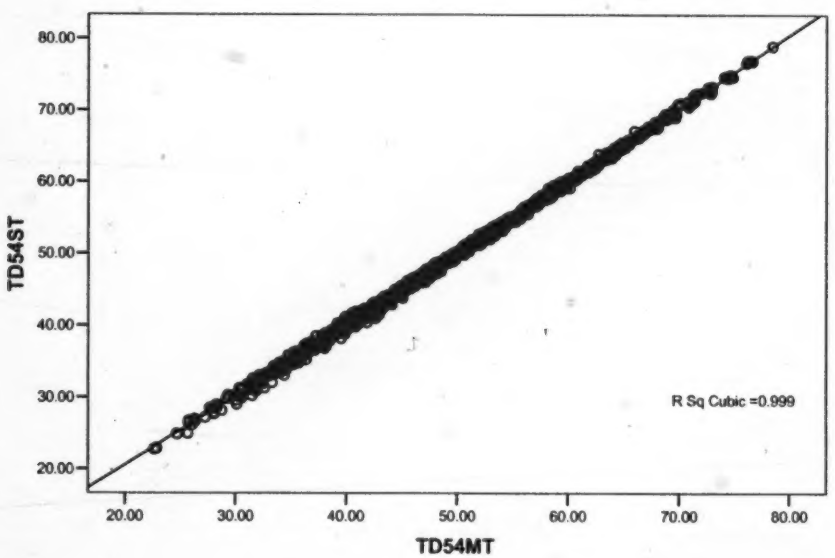


Figure 10: Scatter plot of CTM and MIRM - 0009ST

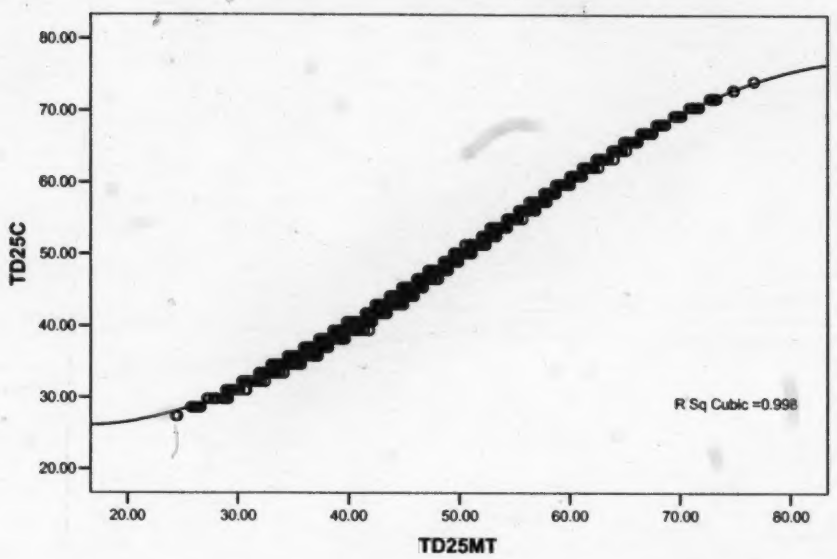


Figure 11: Scatter plot of UIRM (T) and MIRM - 0009ST

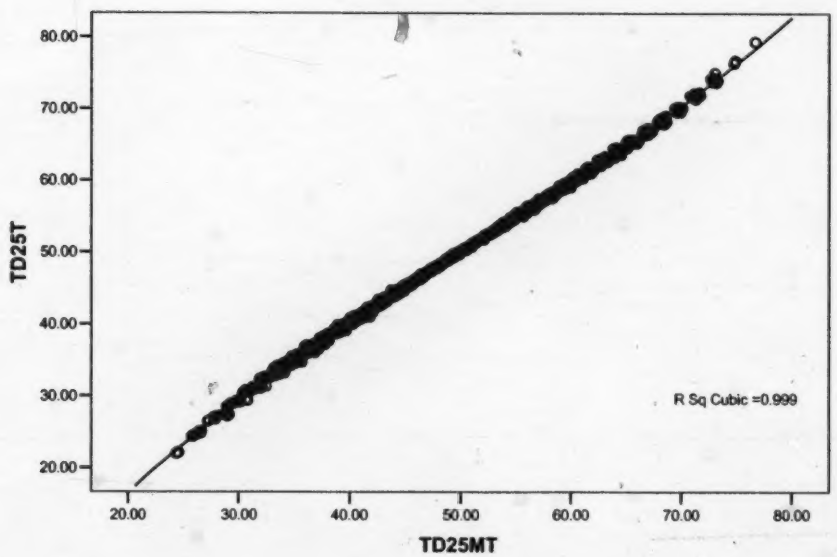


Figure 12: Scatter plot of UIRM (S) and MIRM - 0009ST

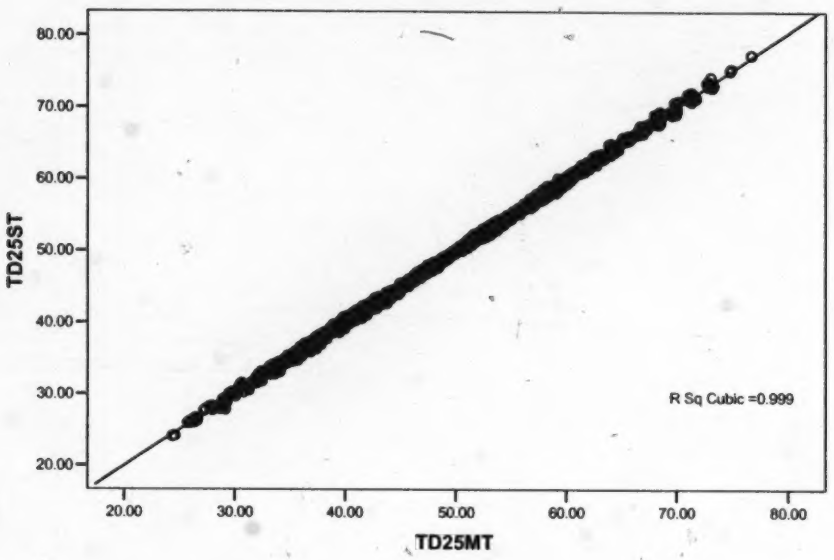


Figure 13: Scatter plot of CTM and MIRM - 0-100ST

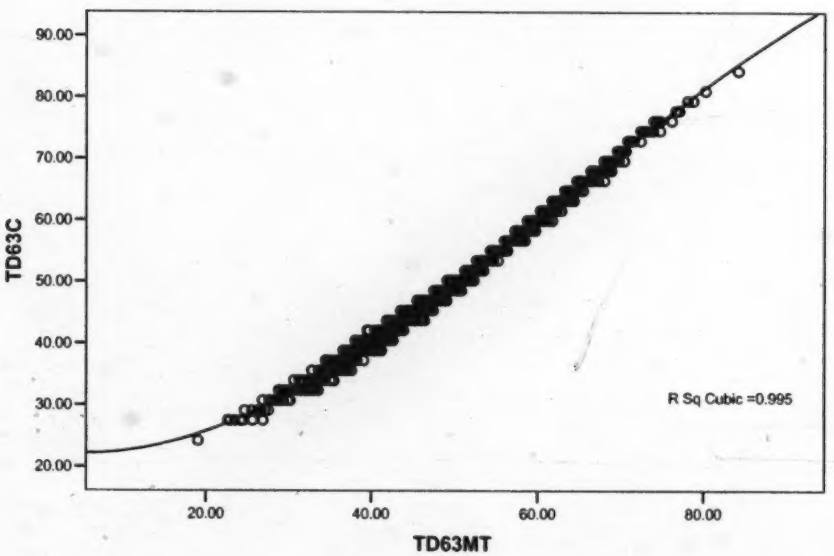


Figure 14: Scatter plot of UIRM (T) and MIRM - 0-100ST

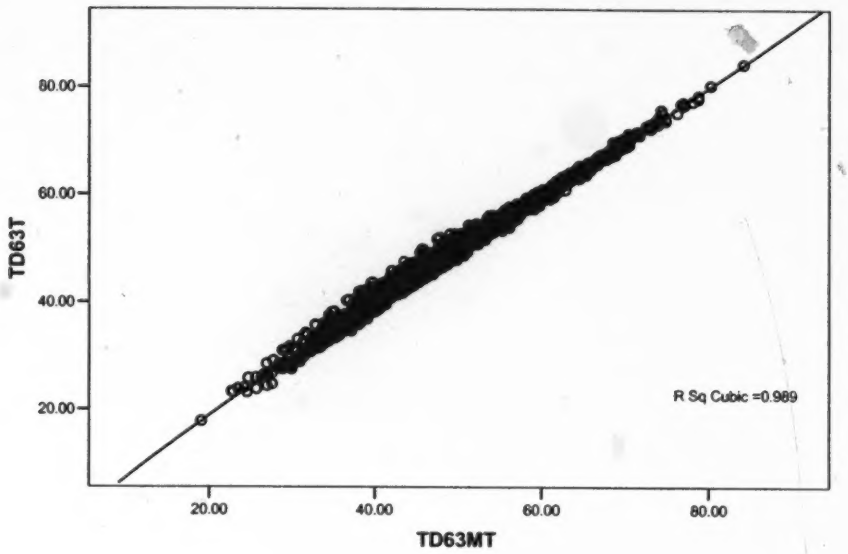


Figure 15: Scatter plot of UIRM (S) and MIRM - 0-100ST

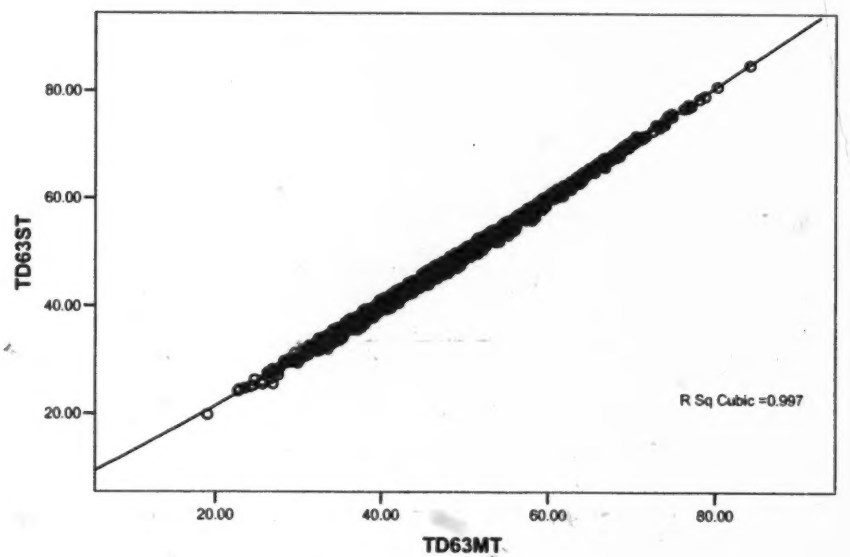


Figure 16: Scatter plot of CTM and MIRM - 0-103ST

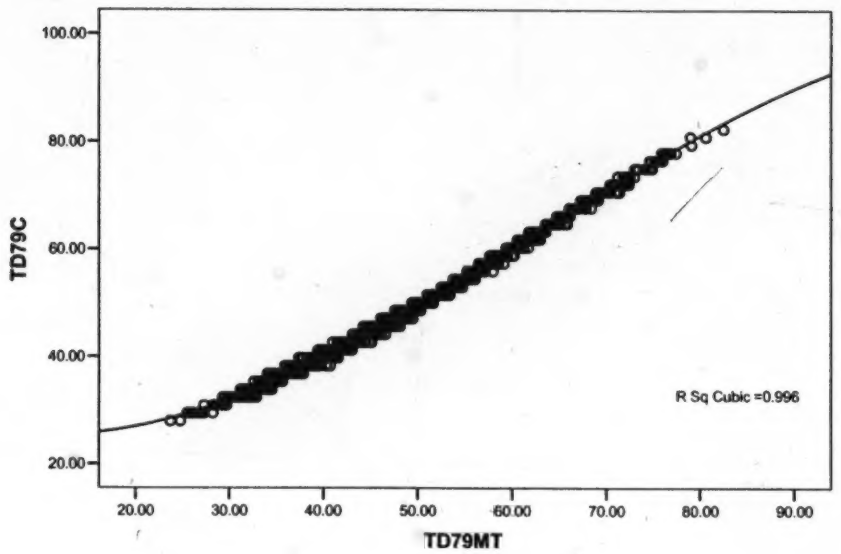


Figure 17: Scatter plot of UIRM (T) and MIRM - 0-103ST

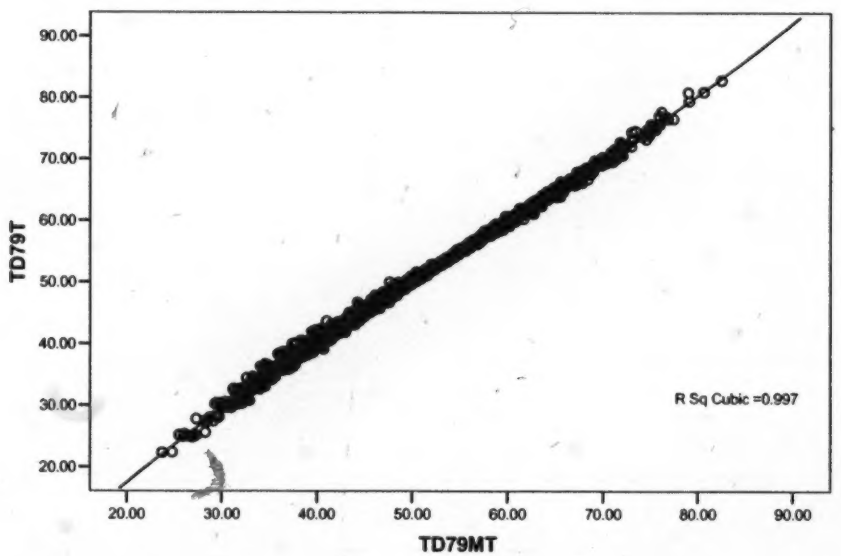


Figure 18: Scatter plot of UIRM (S) and MIRM - 0-103ST

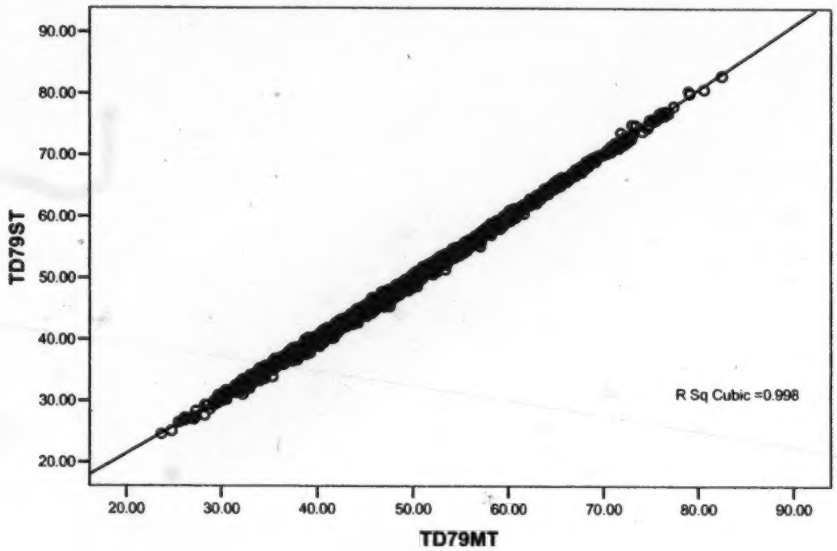


Figure 19: Scatter plot of CTM and MIRM - 0-106ST

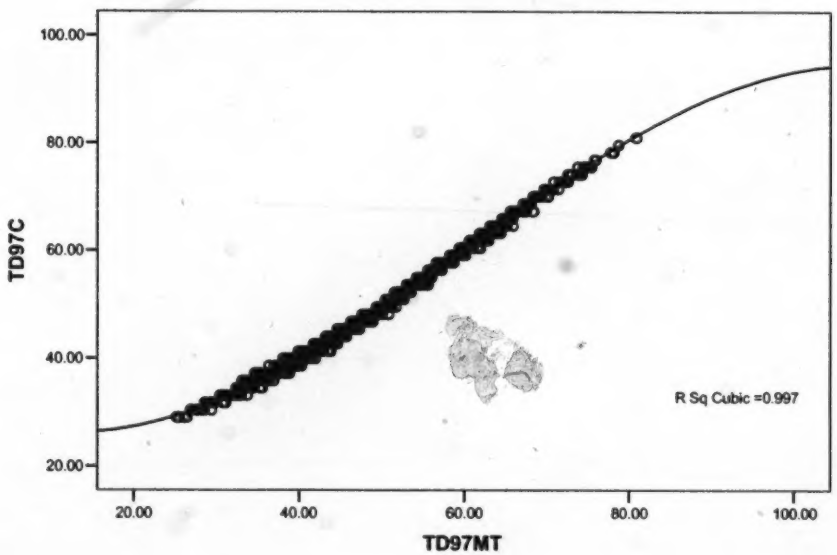


Figure 20: Scatter plot of UIRM (T) and MIRM - 0-106ST

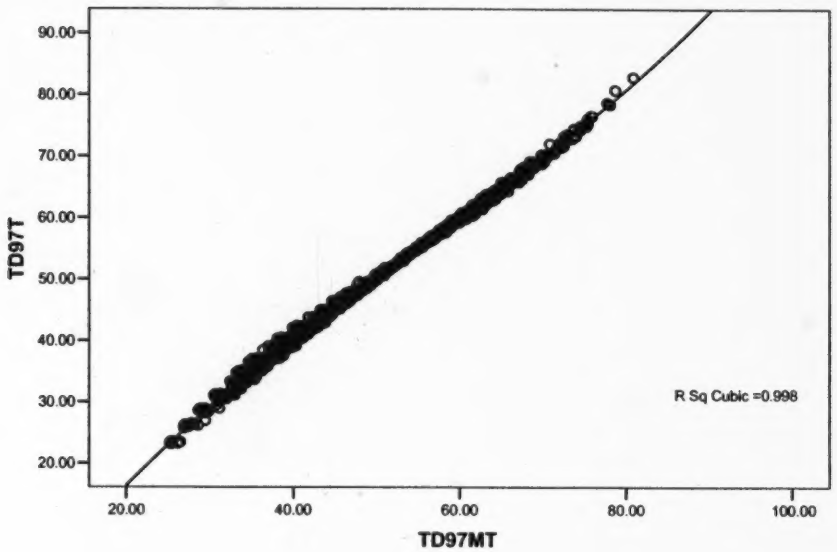


Figure 21: Scatter plot of UIRM (S) and MIRM - 0-106ST

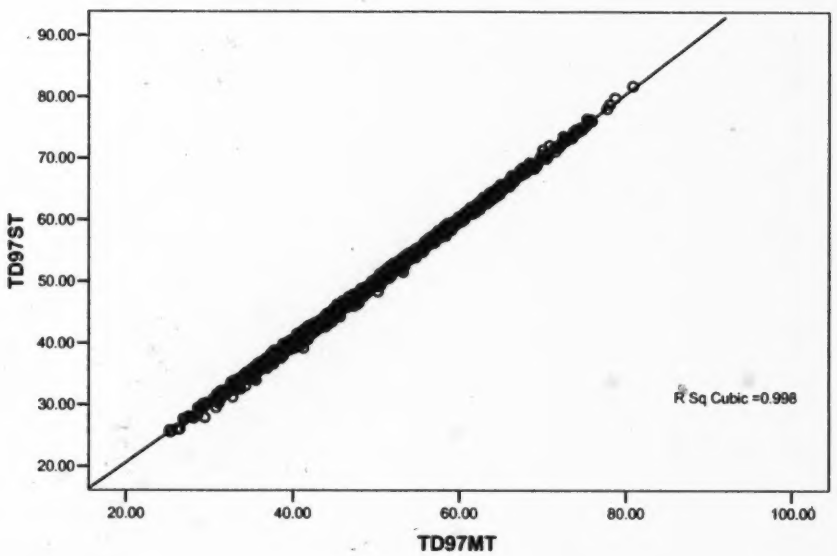


Figure 22: Scatter plot of CTM and MIRM - 0-109ST

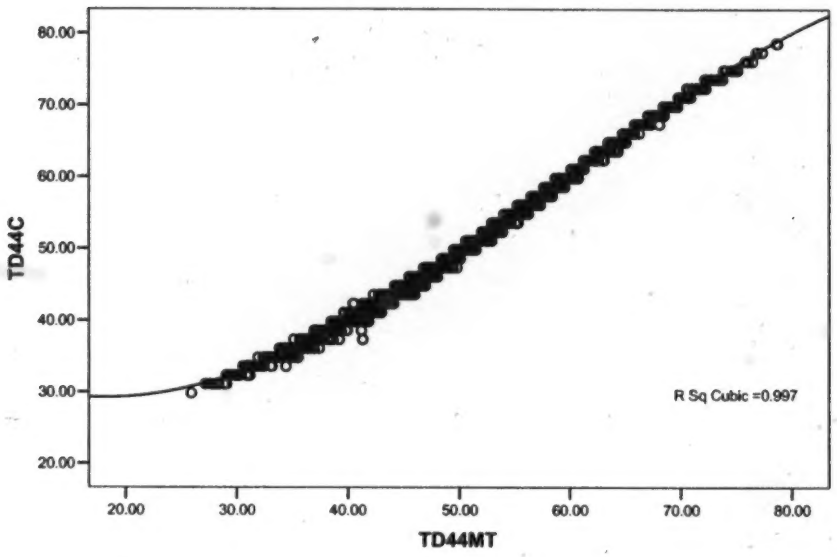


Figure 23: Scatter plot of UIRM (T) and MIRM - 0-109ST

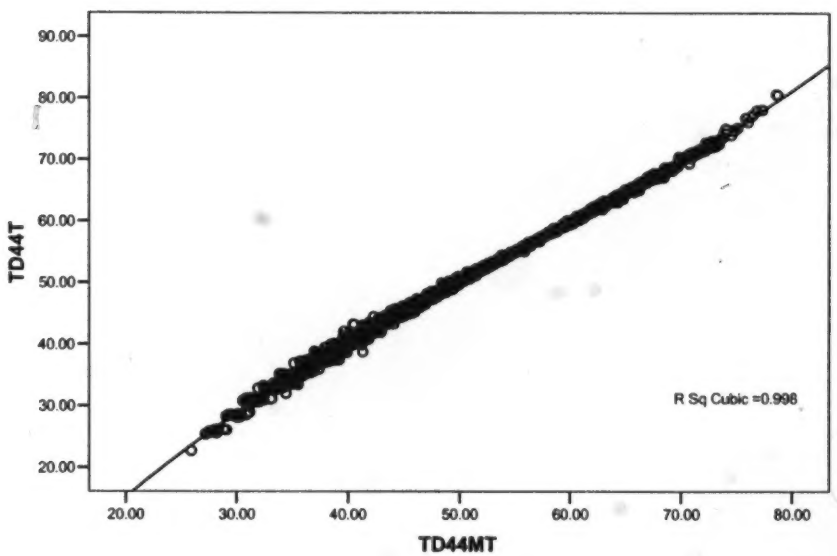


Figure 24: Scatter plot of UIRM (S) and MIRM - 0-109ST

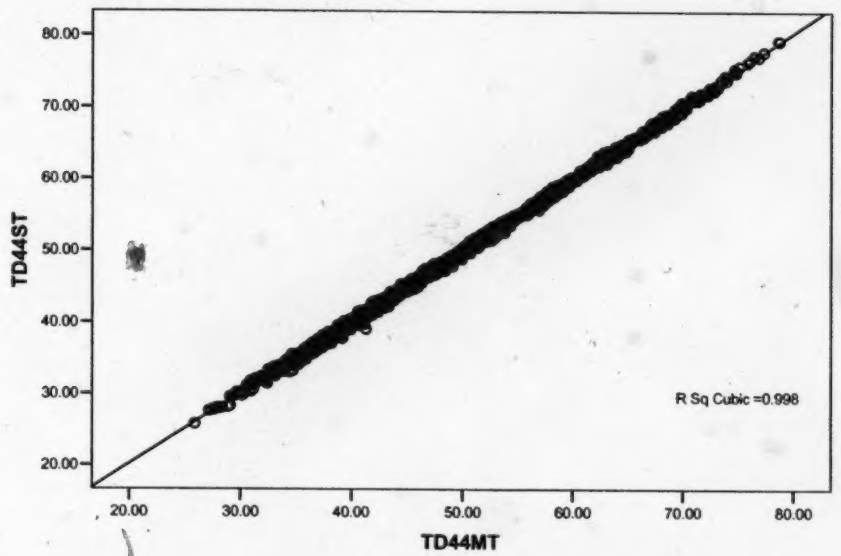


Figure 25: Scatter plot of CTM and MIRM - 0000CT

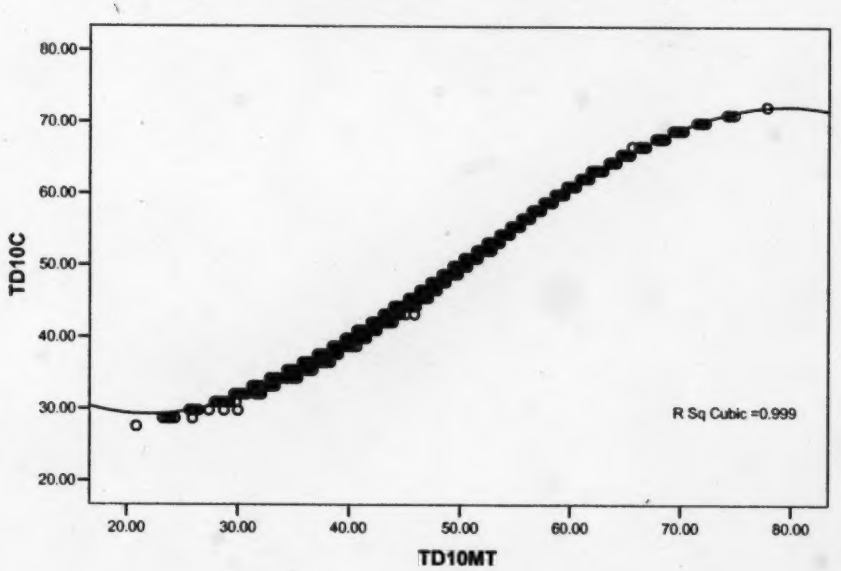


Figure 26: Scatter plot of UIRM (T) and MIRM - 0000CT

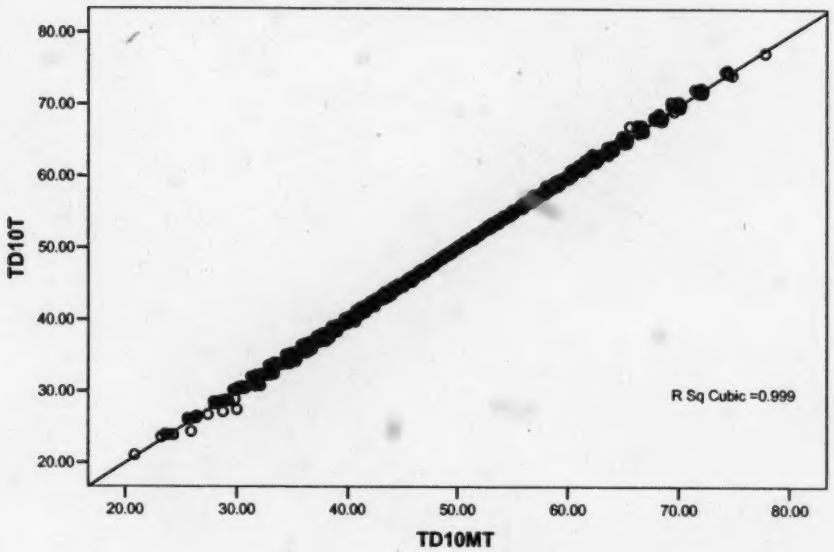


Figure 27: Scatter plot of UIRM (S) and MIRM - 0000CT

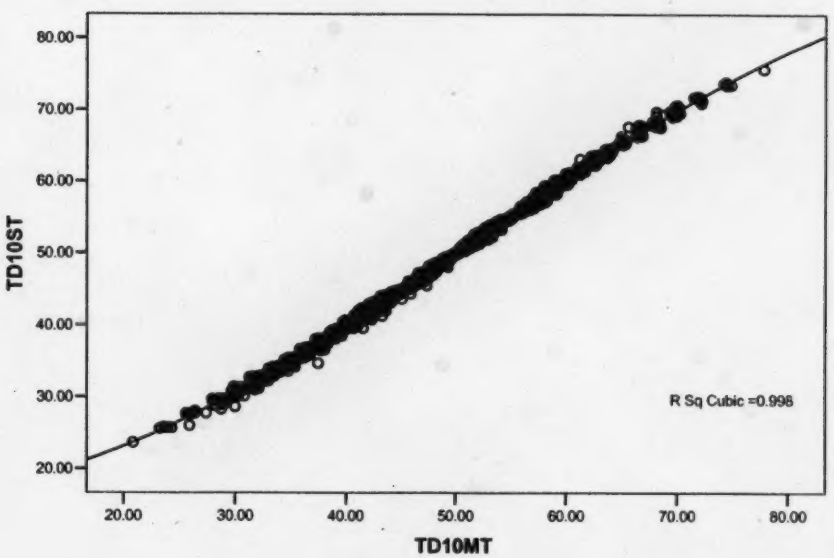


Figure 28: Scatter plot of CTM and MIRM - 0003CT

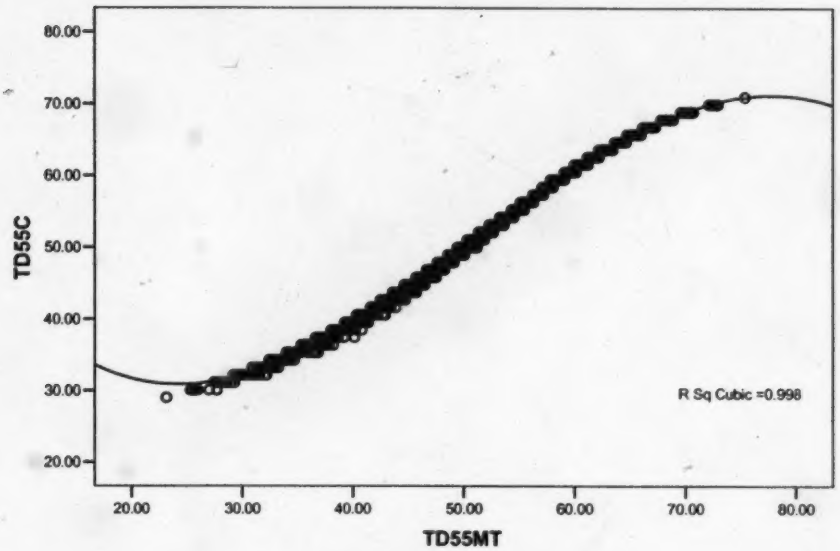


Figure 29: Scatter plot of UIRM (T) and MIRM - 0003CT

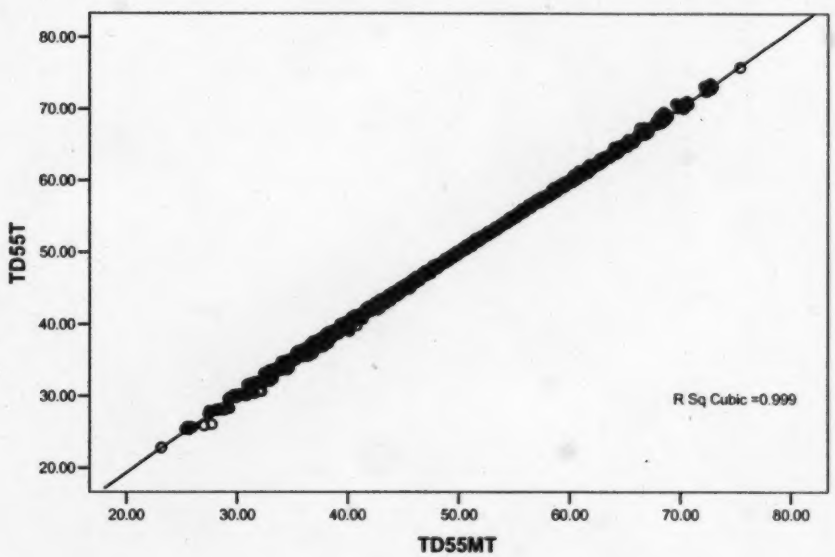


Figure 30: Scatter plot of UIRM (S) and MIRM - 0003CT

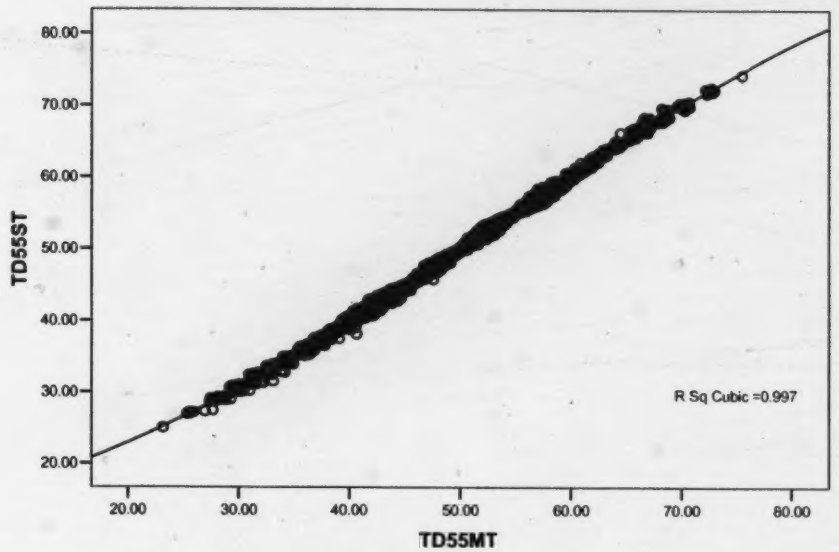


Figure 31: Scatter plot of CTM and MIRM - 0006CT

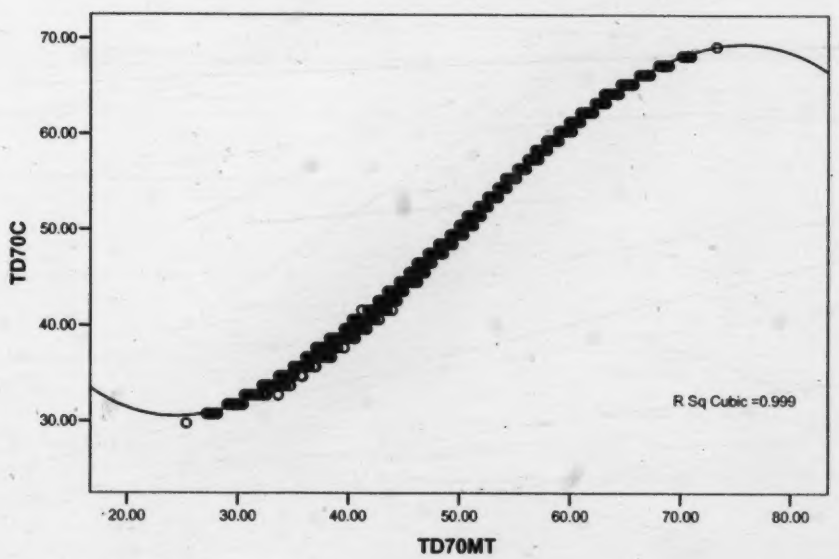


Figure 32: Scatter plot of UIRM (T) and MIRM - 0006CT

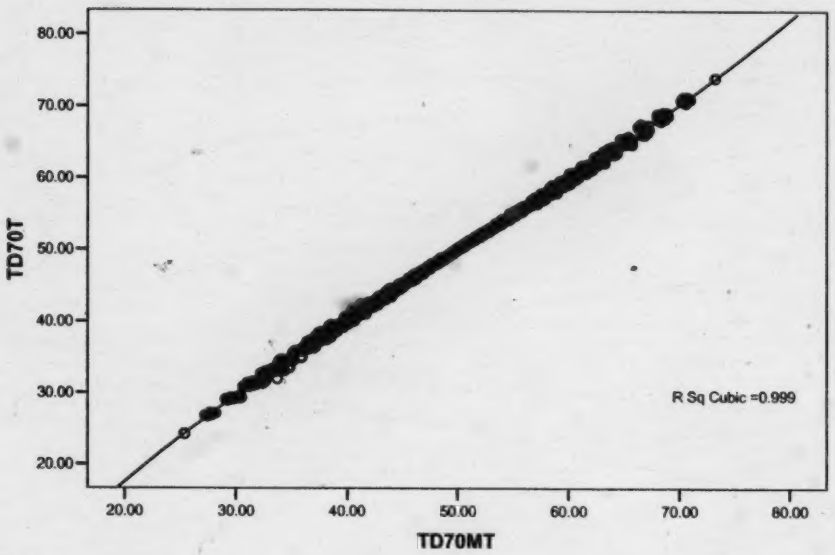


Figure 33: Scatter plot of UIRM (S) and MIRM - 0006CT

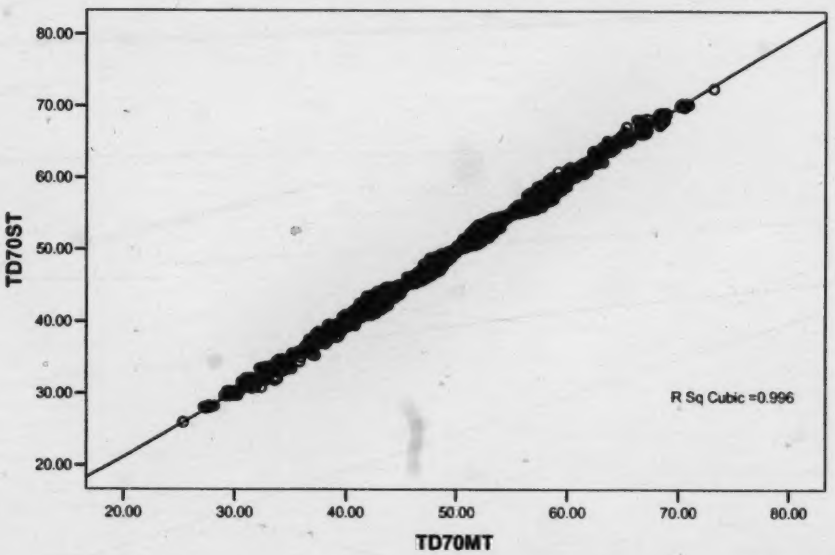


Figure 34: Scatter plot of CTM and MIRM - 0-100CT

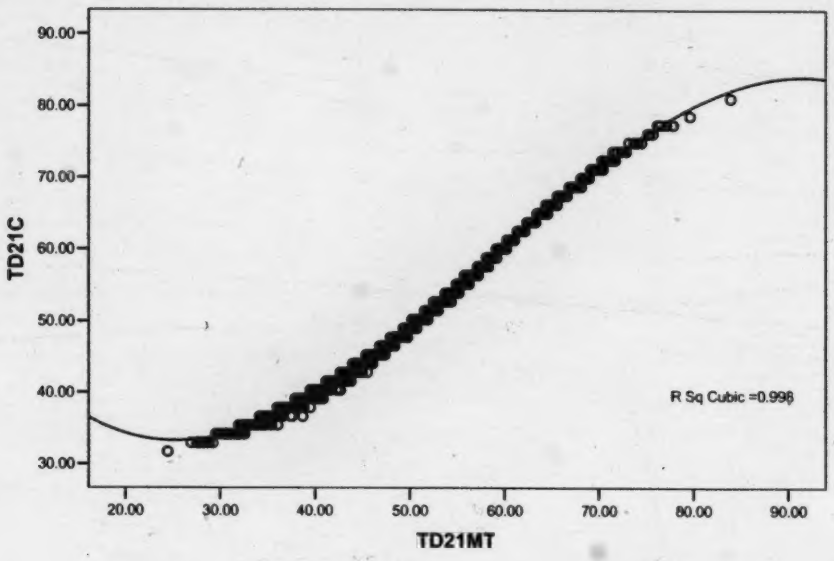


Figure 35: Scatter plot of UIRM (T) and MIRM - 0-100CT

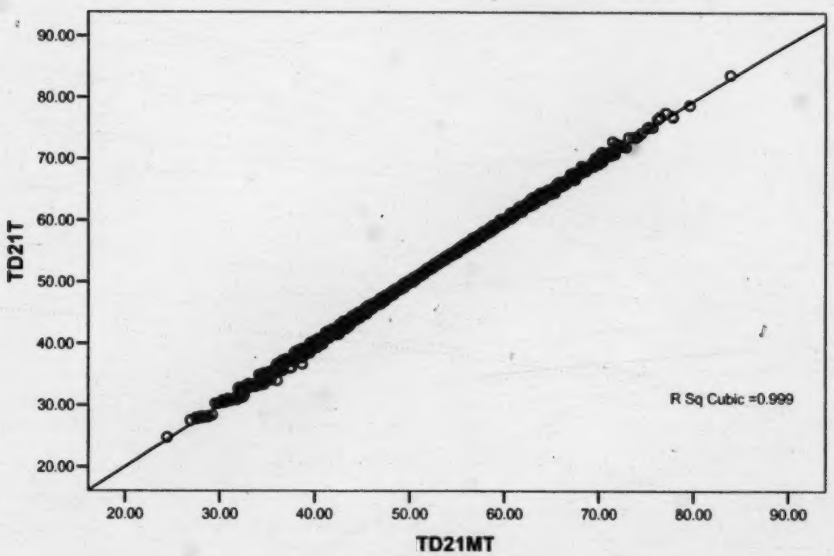


Figure 36: Scatter plot of UIRM (S) and MIRM - 0-100CT

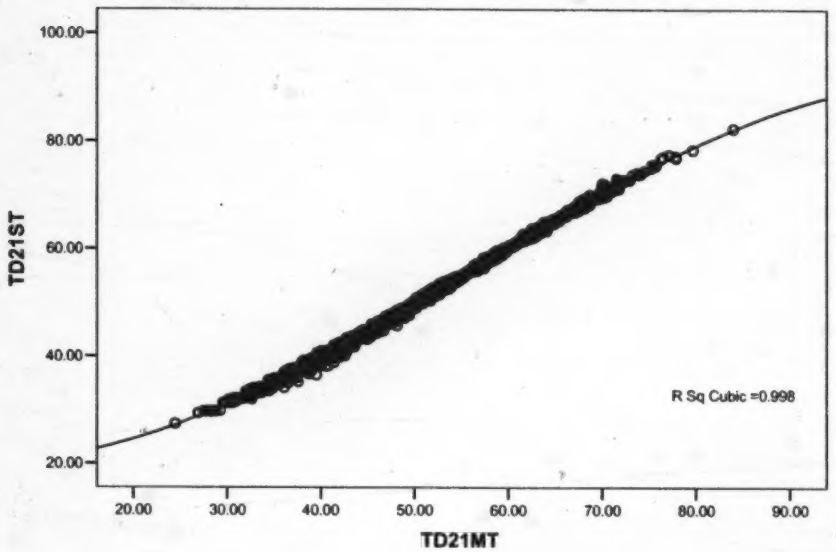


Figure 37: Scatter plot of CTM and MIRM - 0-103CT

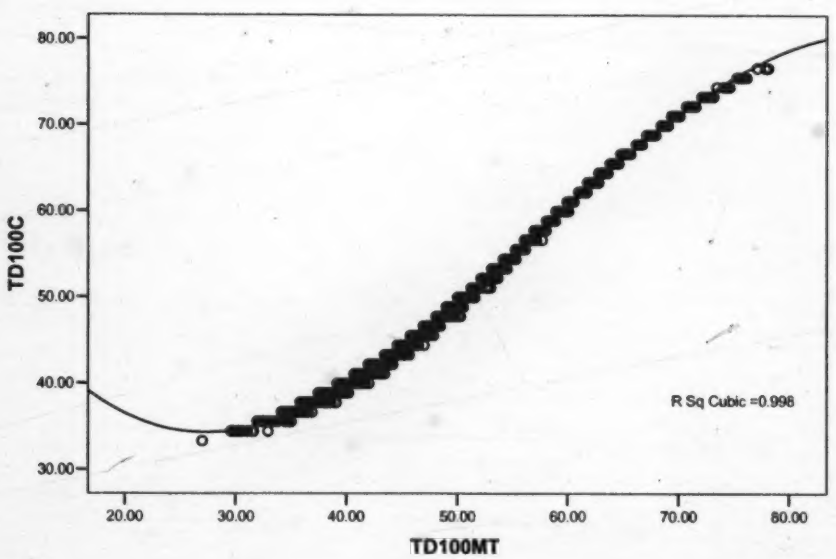


Figure 38: Scatter plot of UIRM (T) and MIRM - 0-103CT

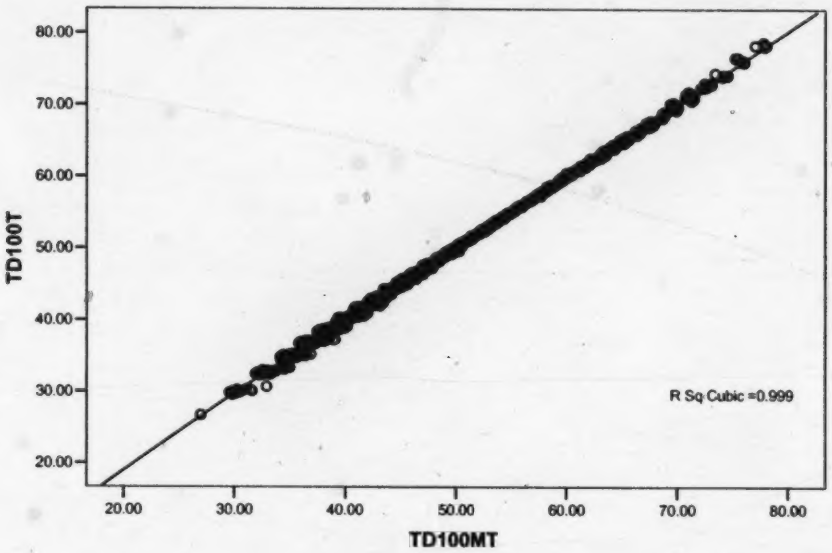


Figure 39: Scatter plot of UIRM (S) and MIRM - 0-103CT

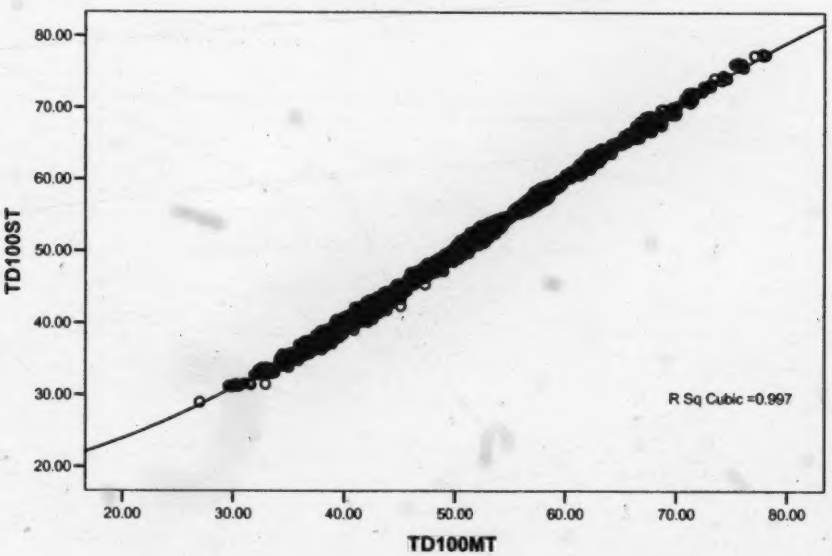


Figure 40: Scatter plot of CTM and MIRM - 0-106CT

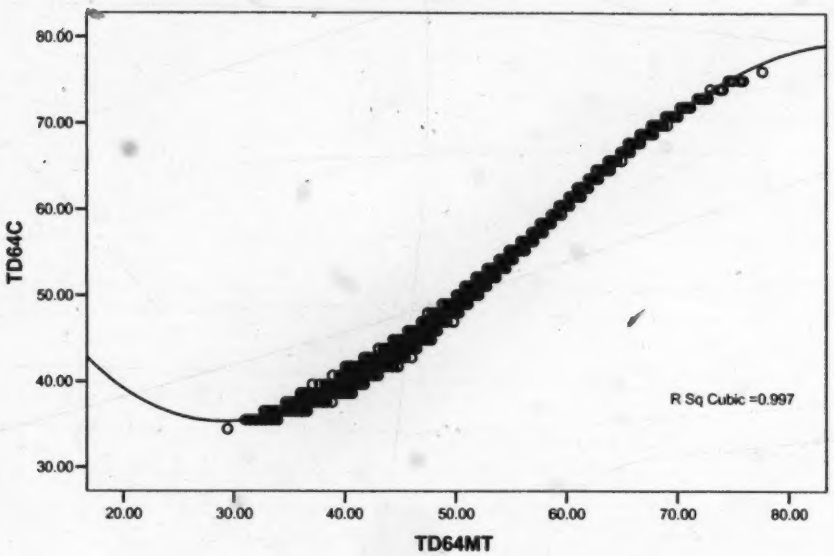


Figure 41: Scatter plot of UIRM (T) and MIRM - 0-106CT

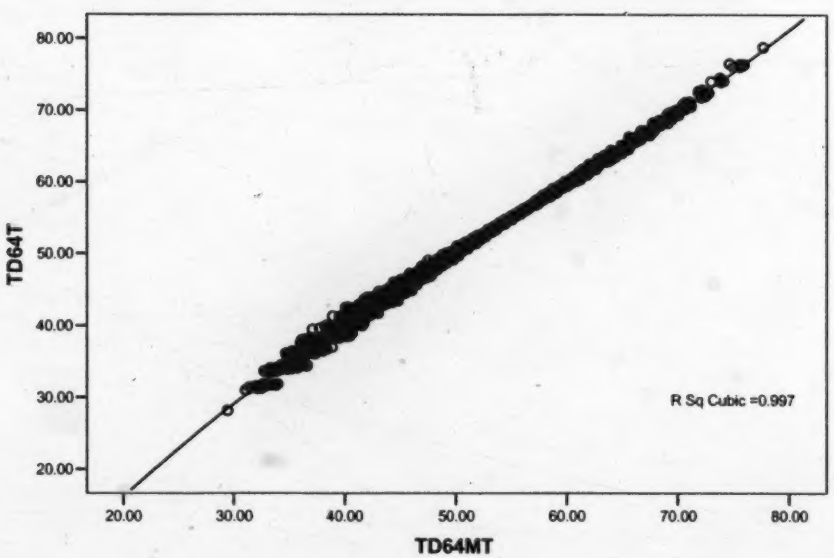
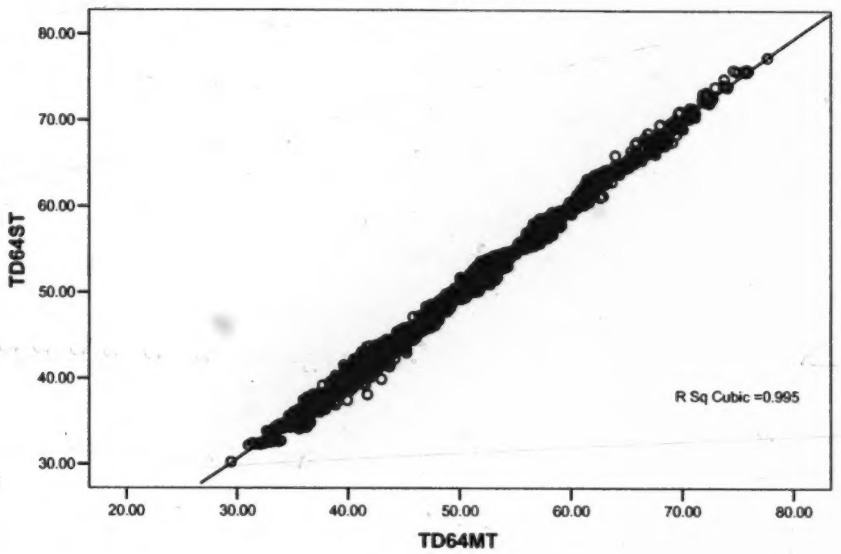


Figure 42: Scatter plot of UIRM (S) and MIRM - 0-106CT



Appendix B

Subtest Scores

This appendix provides the scatter plots of the subtest scores yielded by the MIRM and each of the CTM and UIRM (S). For each condition, the same randomly selected data set used for the total score was also used for the subtest scores. Figure 1 represents the scatter plot of the subtest scores of dimension 1 of the CTM – MIRM comparison for condition 0000S where the scores from the 9th replication were used. Figure 1a represents the scatter plot of subtest scores for dimension 2.

Figure 1: Scatter plot of CTM and MIRM - Subtest 1 - 0000SS

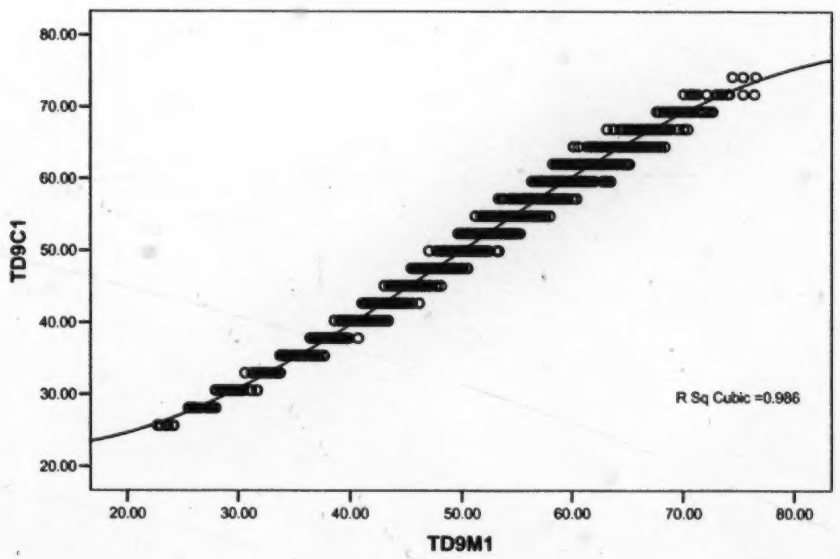


Figure 1a: Scatter plot of CTM and MIRM - Subtest 2 - 0000SS

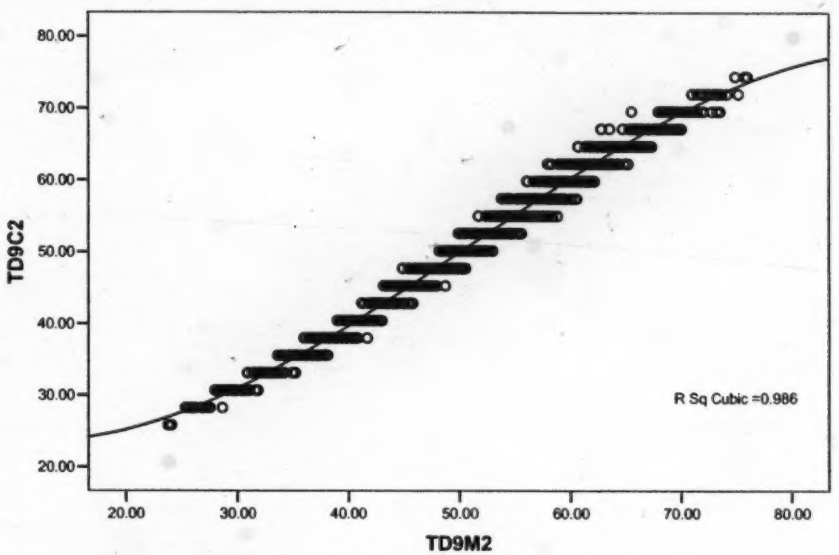


Figure 2: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0000SS

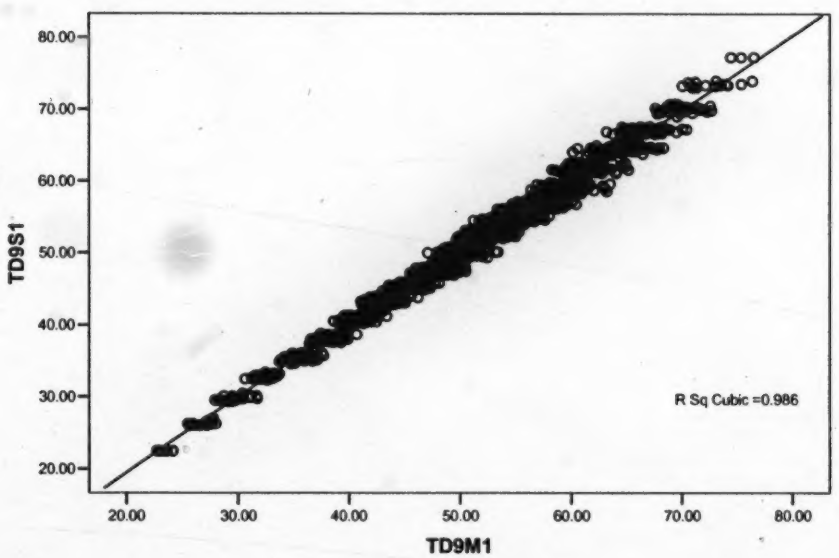


Figure 2a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0000SS

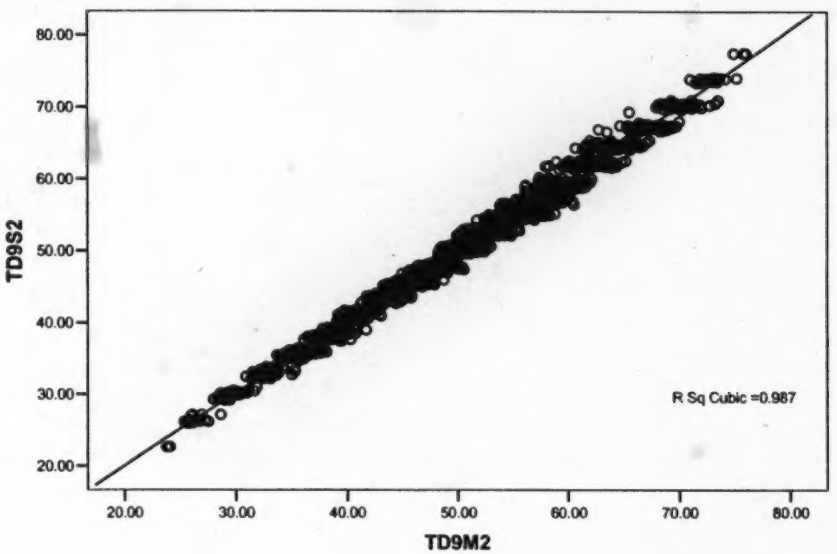


Figure 3: Scatter plot of CTM and MIRM - Subtest 1 - 0003SS

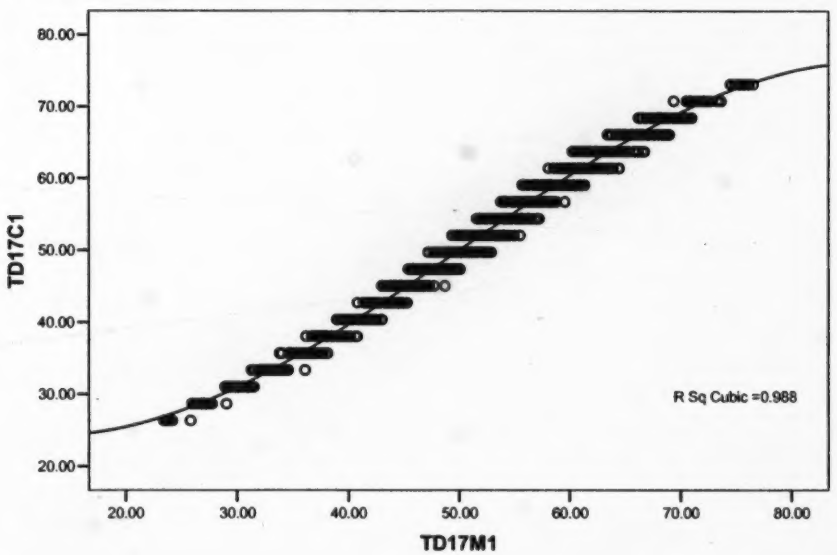


Figure 3a: Scatter plot of CTM and MIRM - Subtest 2 - 0003SS

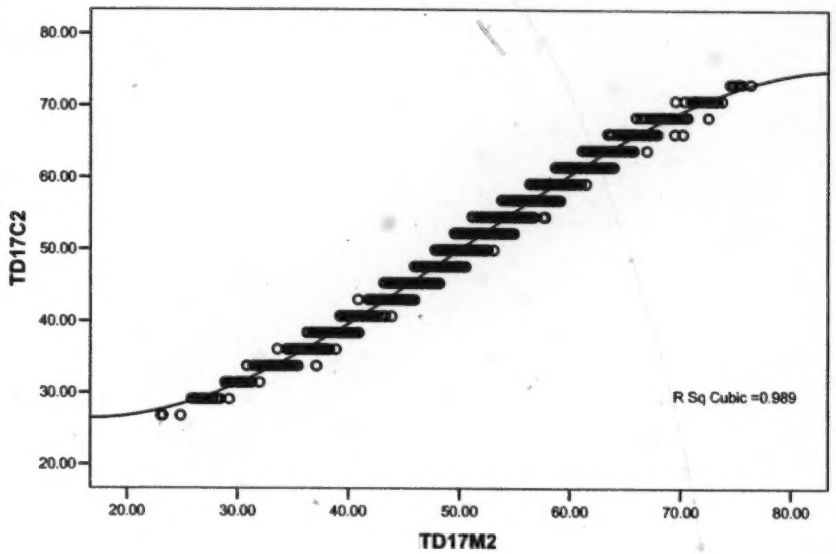


Figure 4: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0003SS

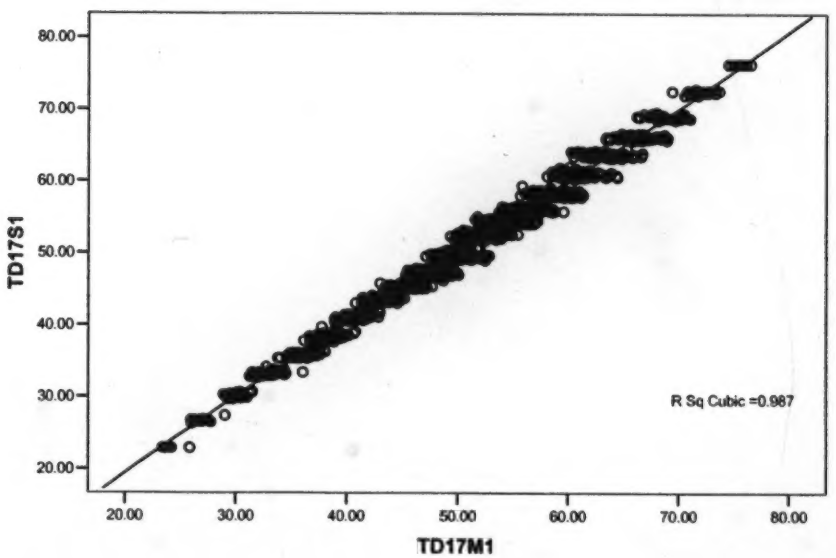


Figure 4a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0003SS

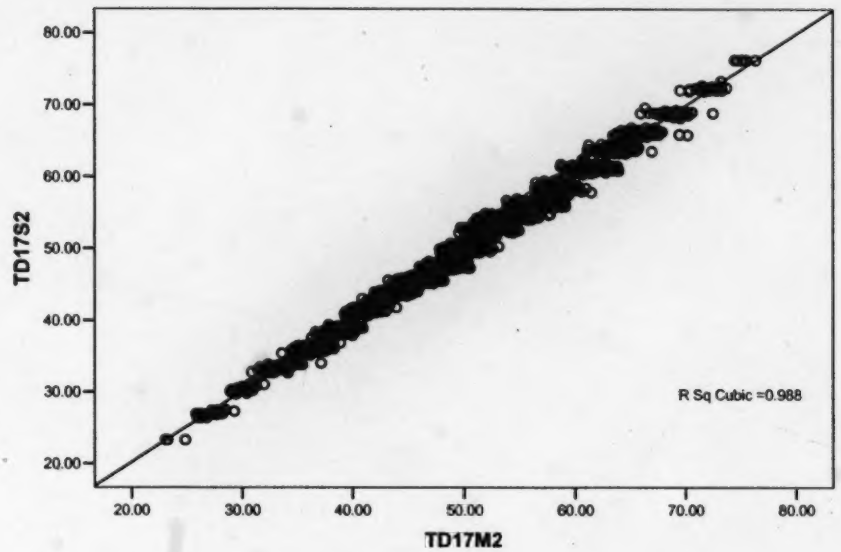


Figure 5: Scatter plot of CTM and MIRM - Subtest 1 - 0006SS

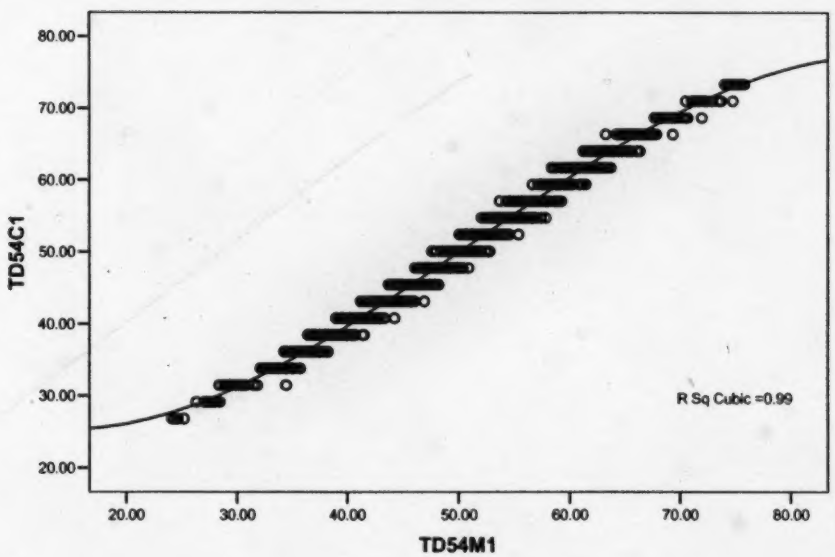


Figure 5a: Scatter plot of CTM and MIRM - Subtest 2 - 0006SS

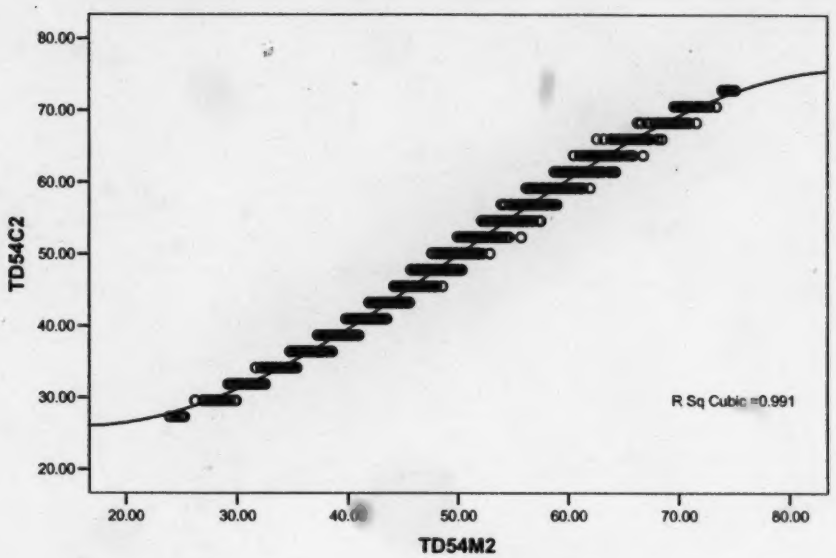


Figure 6: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0006SS

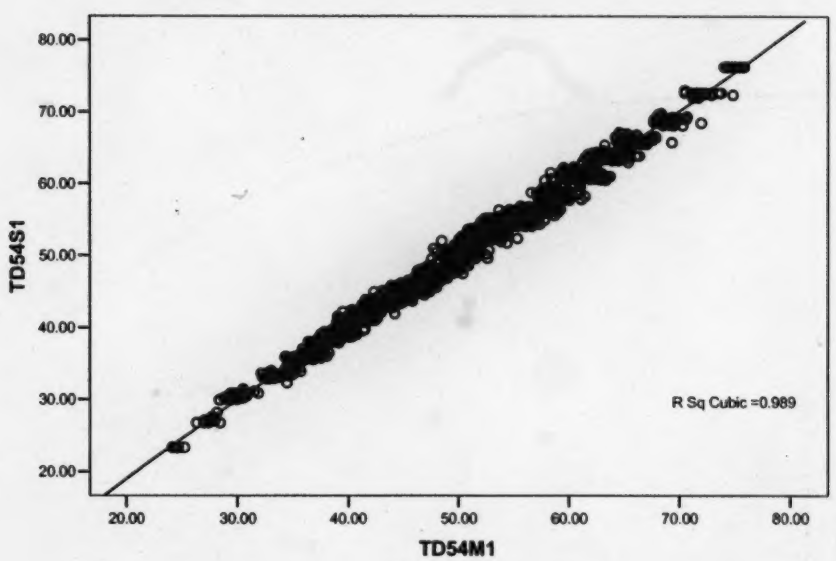


Figure 6a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0006SS

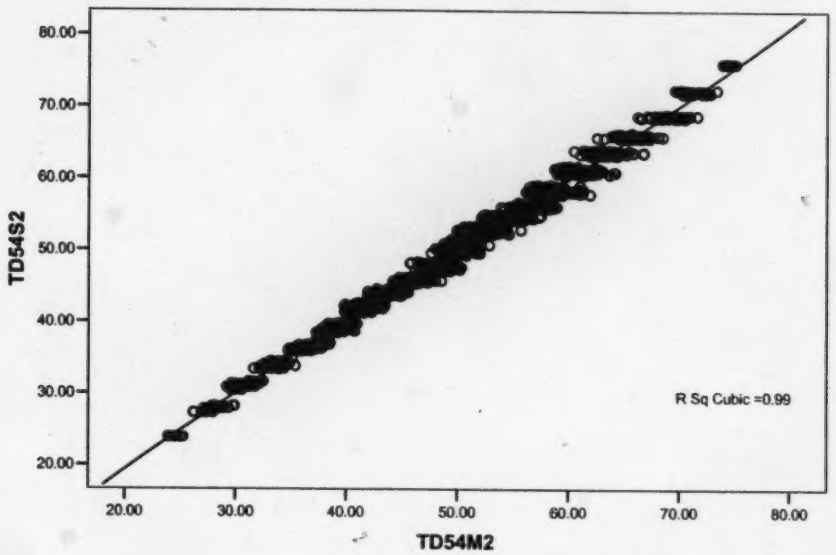


Figure 7: Scatter plot of CTM and MIRM - Subtest 1 - 0009SS

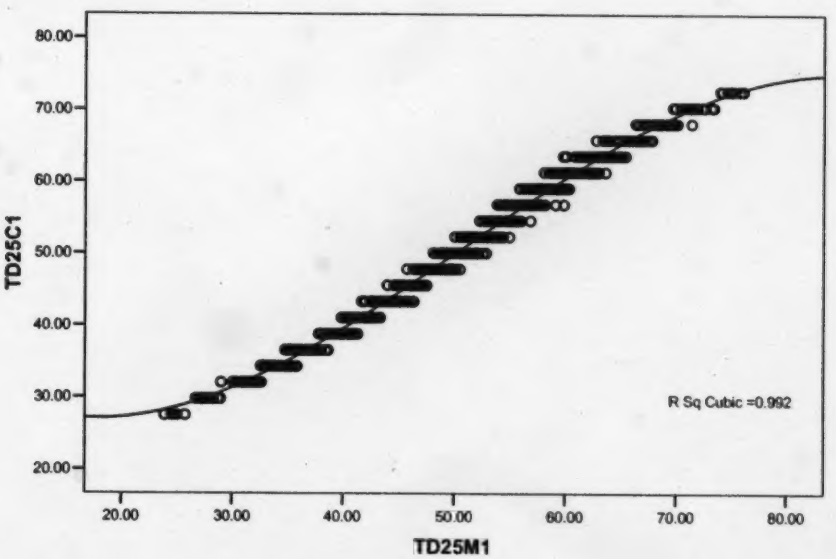


Figure 7a: Scatter plot of CTM and MIRM - Subtest 2 - 0009SS

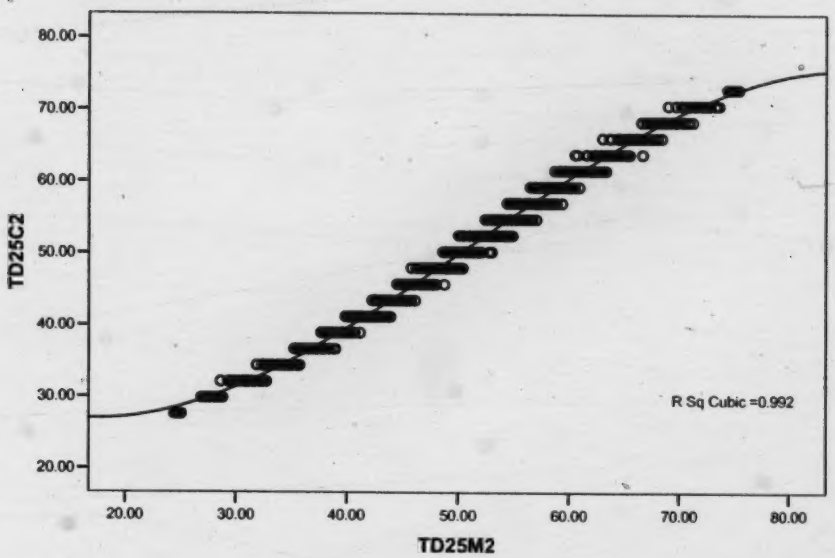


Figure 8: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0009SS

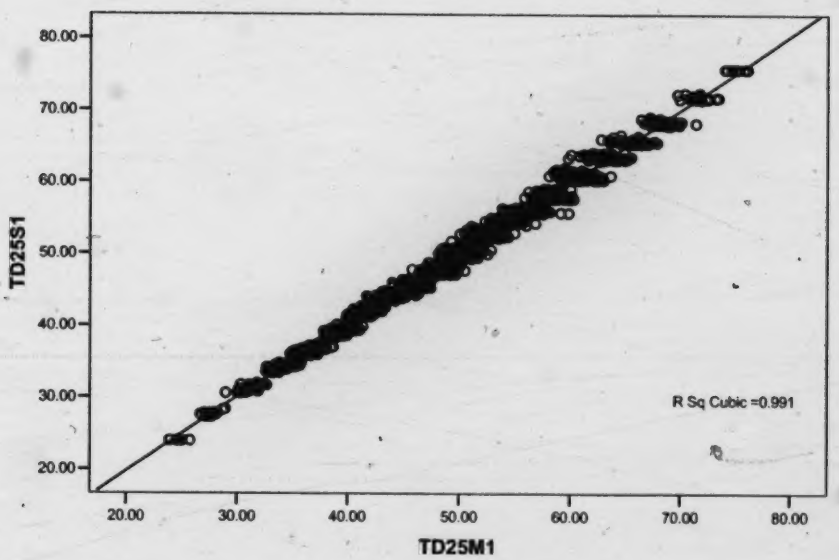


Figure 8a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0009SS

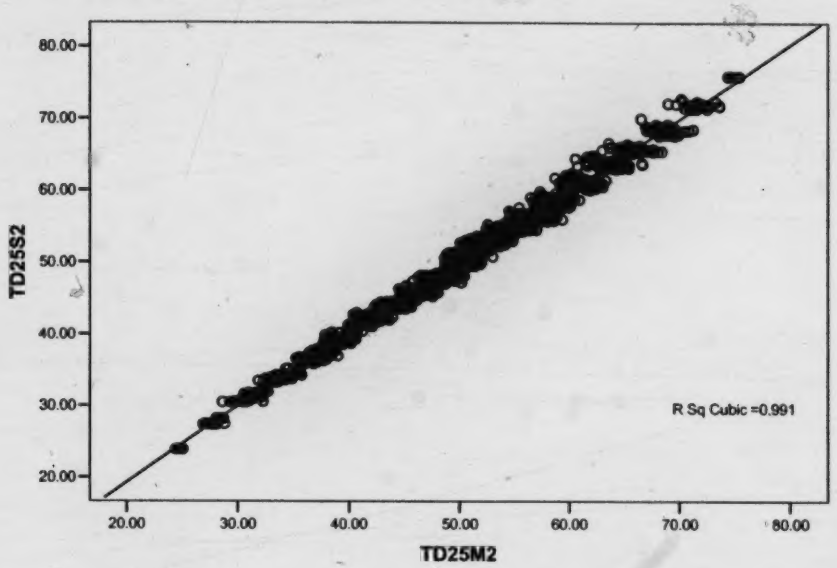


Figure 9: Scatter plot of CTM and MIRM - Subtest 1 - 0-100SS

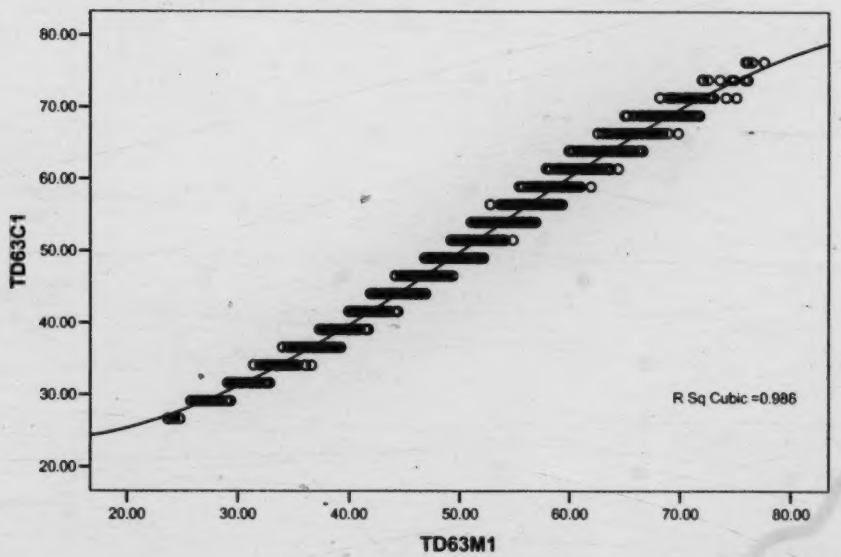


Figure 9a: Scatter plot of CTM and MIRM - Subtest 2 - 0-100SS

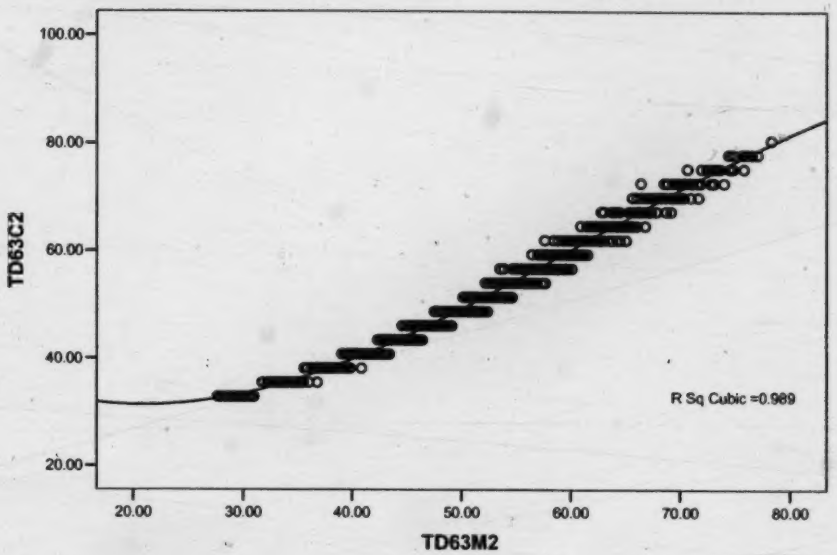


Figure 10: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-100SS

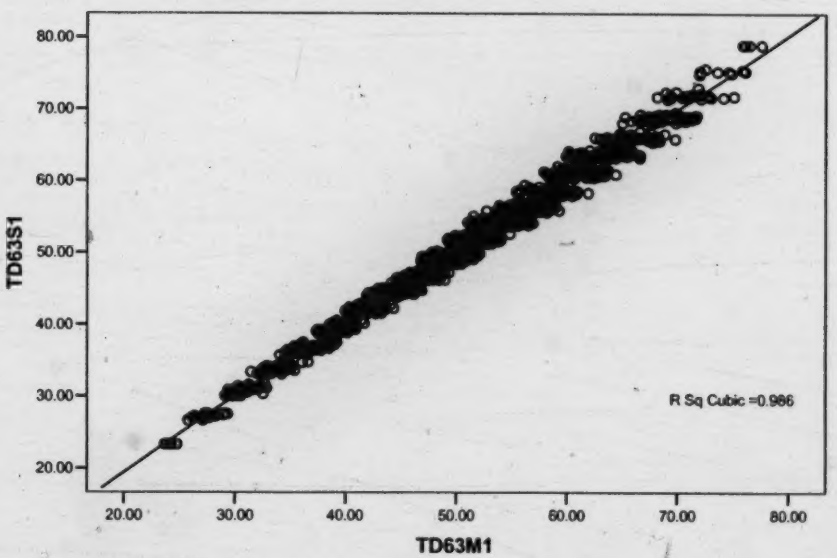


Figure 10a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-100SS

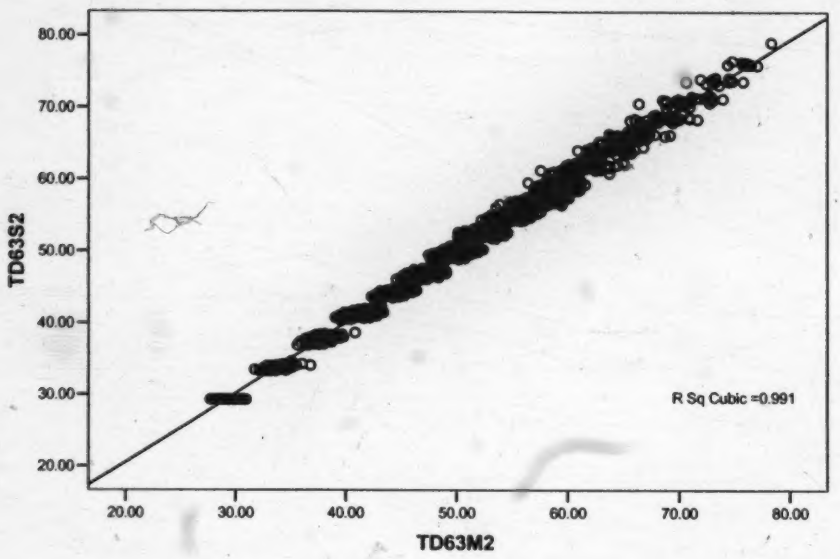


Figure 11: Scatter plot of CTM and MIRM - Subtest 1 - 0-103SS

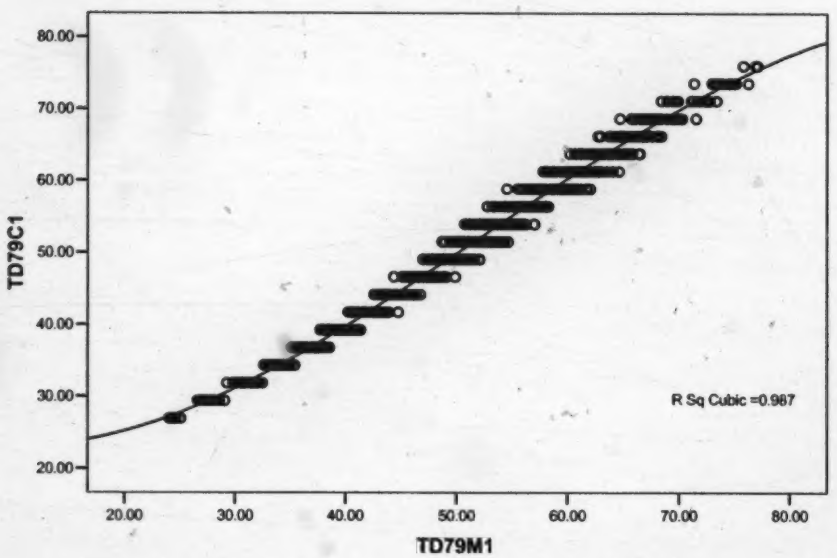


Figure 11a: Scatter plot of CTM and MIRM - Subtest 2 - 0-103SS

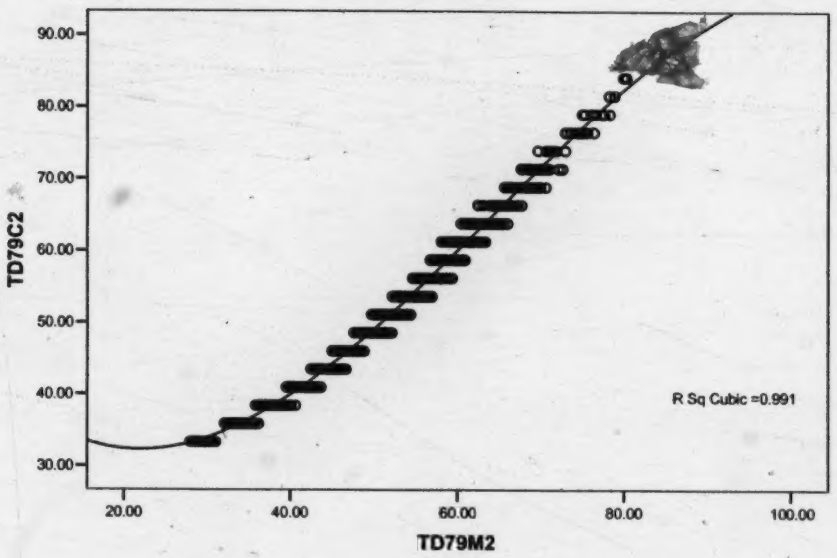


Figure 12: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-103SS

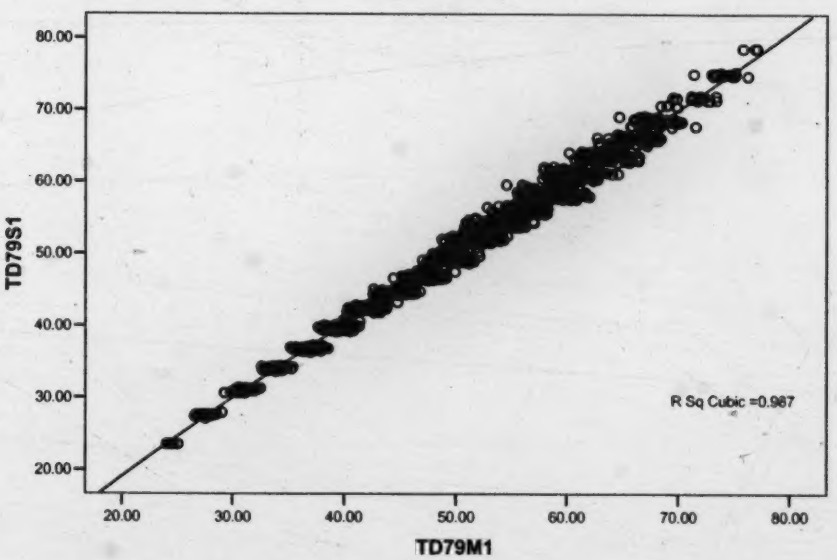


Figure 12a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-103SS

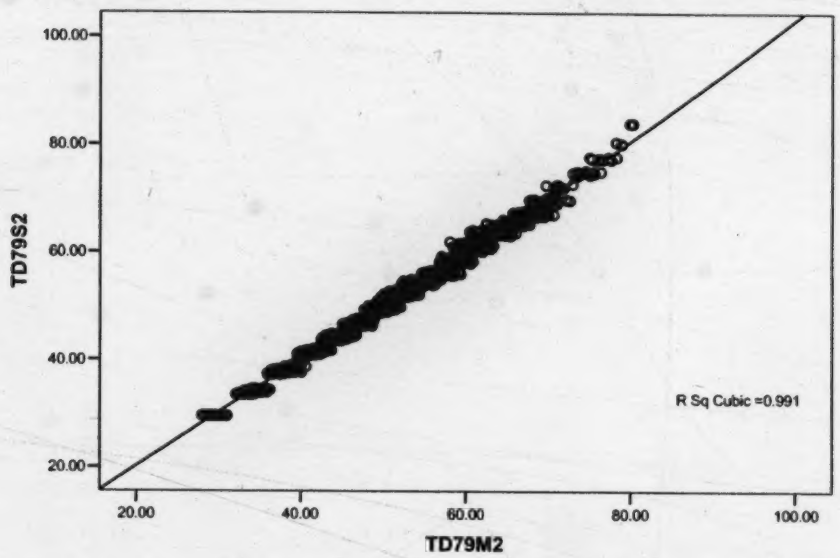


Figure 13: Scatter plot of CTM and MIRM - Subtest 1 - 0-106SS

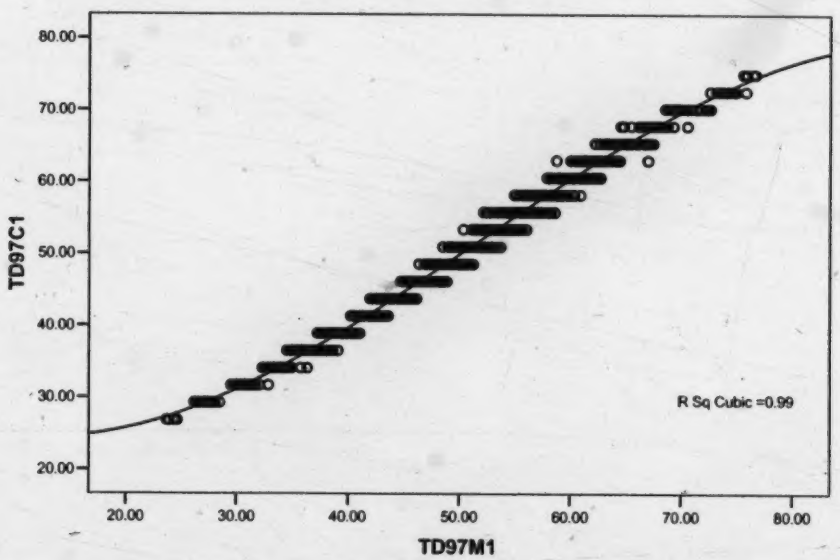


Figure 13a: Scatter plot of CTM and MIRM - Subtest 2 - 0-106SS

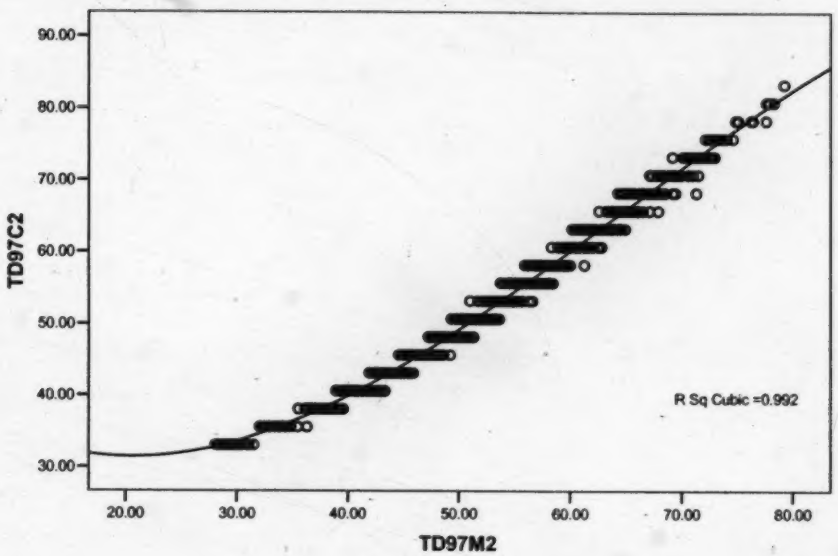


Figure 14: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-106SS

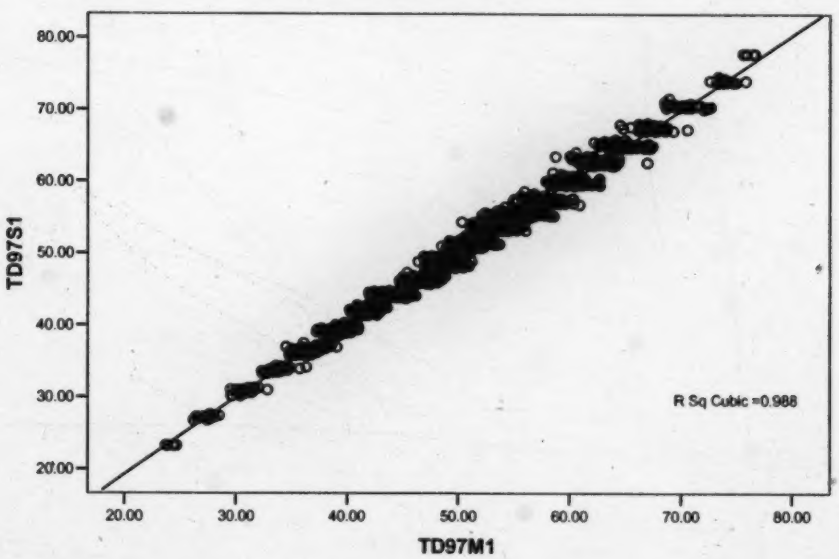


Figure 14a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-106SS

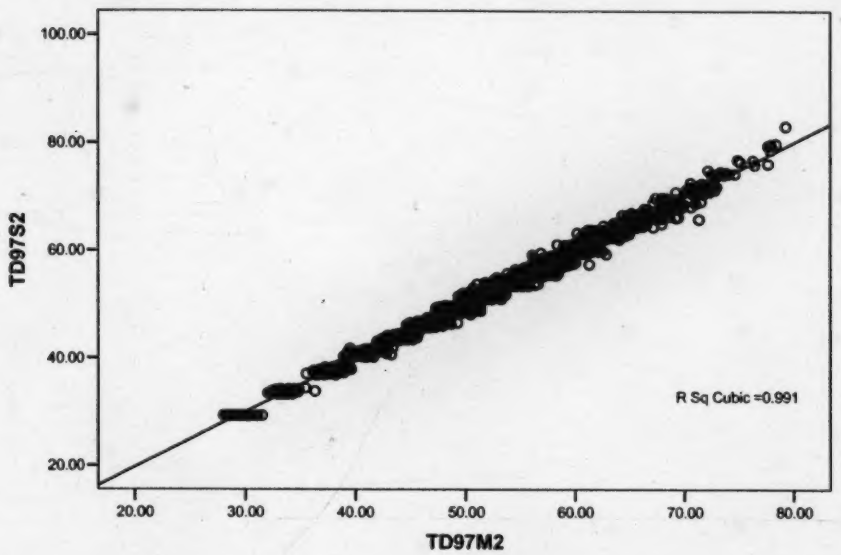


Figure 15: Scatter plot of CTM and MIRM - Subtest 1 - 0-109SS

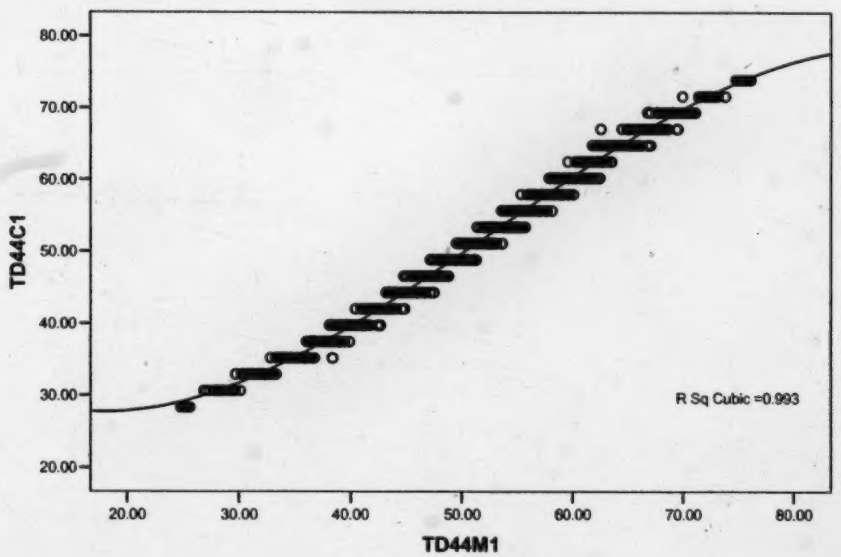


Figure 15a: Scatter plot of CTM and MIRM - Subtest 2 - 0-109SS

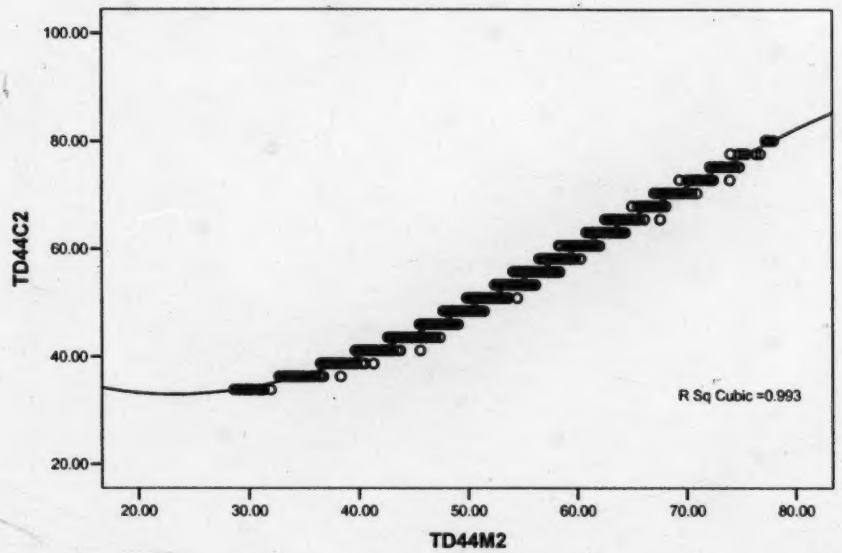


Figure 16: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-109SS

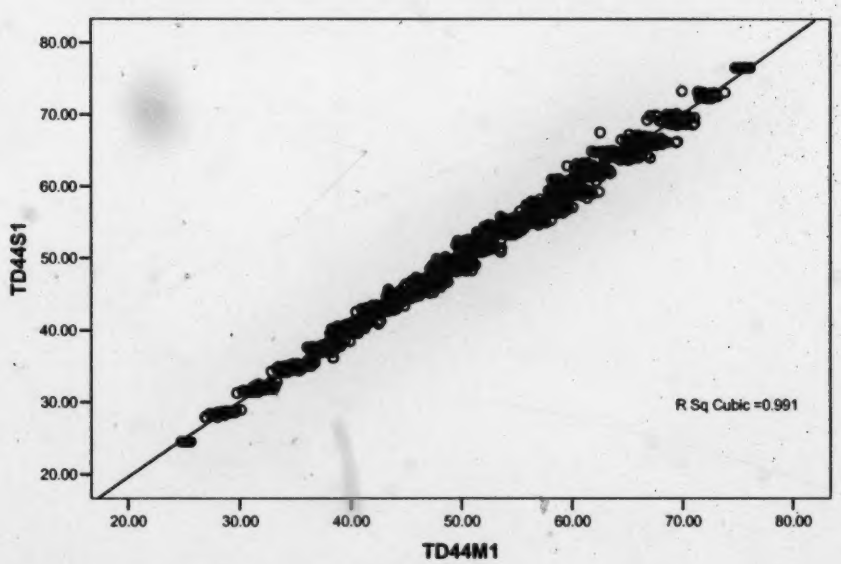


Figure 16a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-109SS

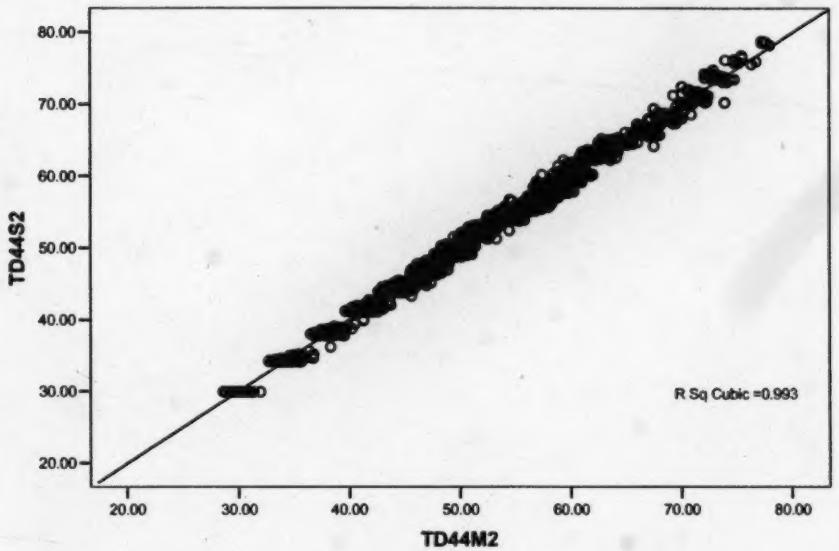


Figure 17: Scatter plot of CTM and MIRM - Subtest 1 - 0000CS

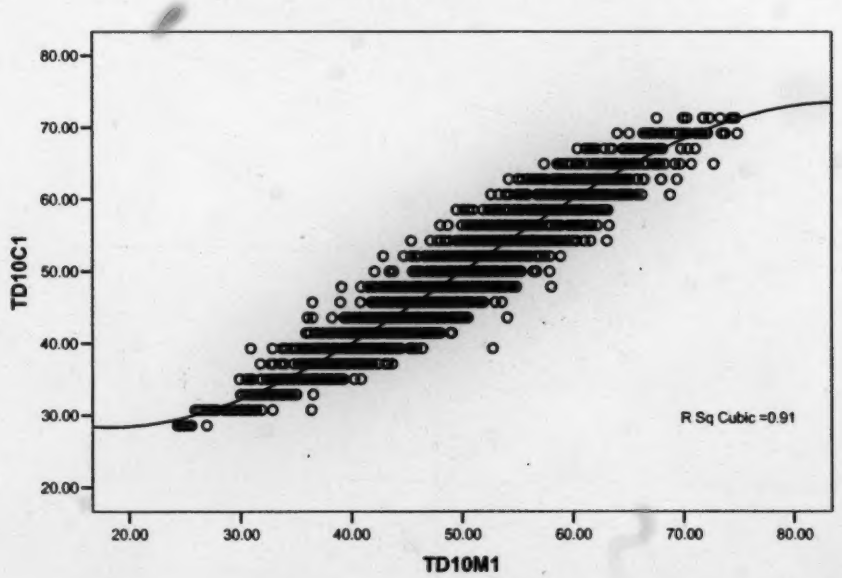


Figure 17a: Scatter plot of CTM and MIRM - Subtest 2 - 0000CS

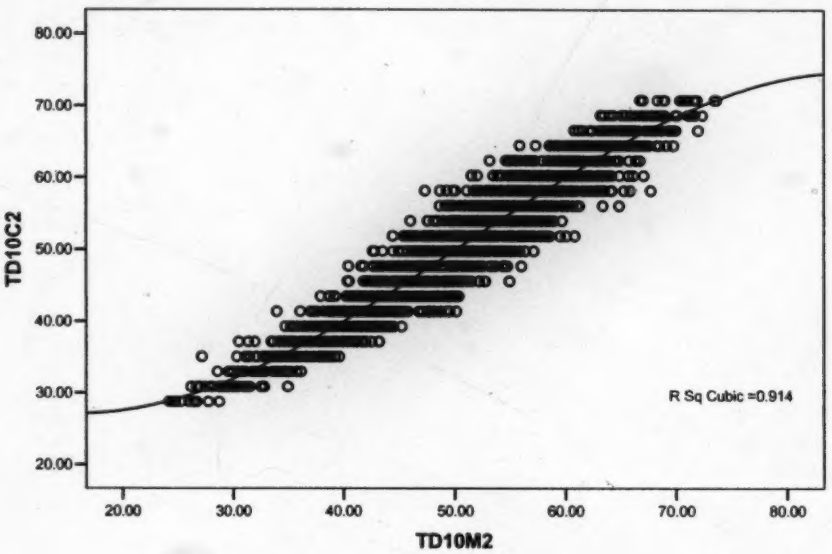


Figure 18: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0000CS

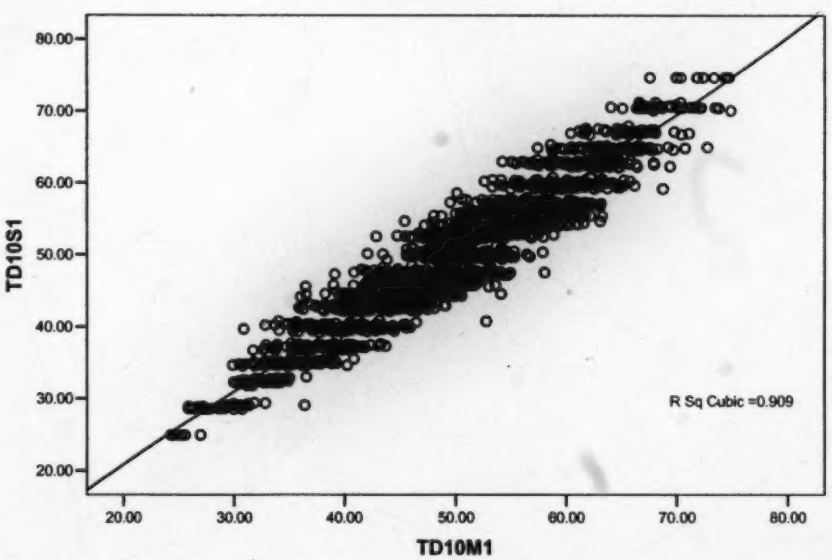


Figure 18a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0000CS

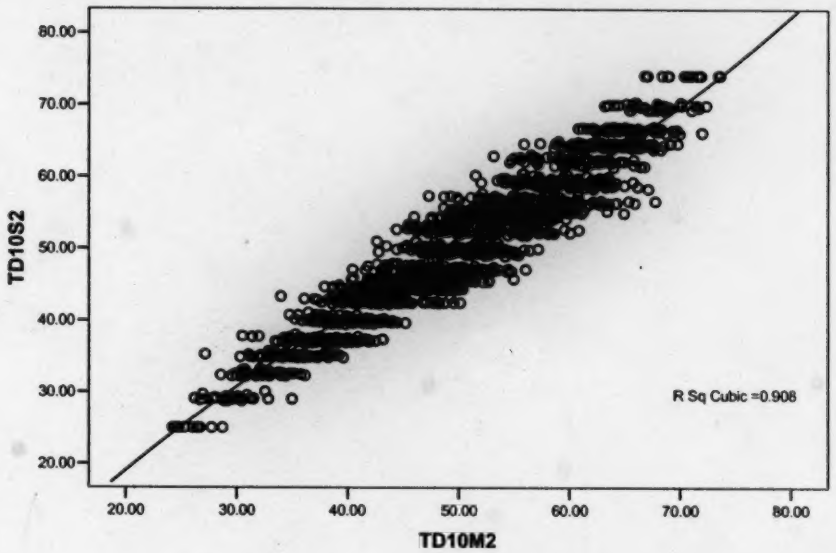


Figure 19: Scatter plot of CTM and MIRM - Subtest 1 - 0003CS

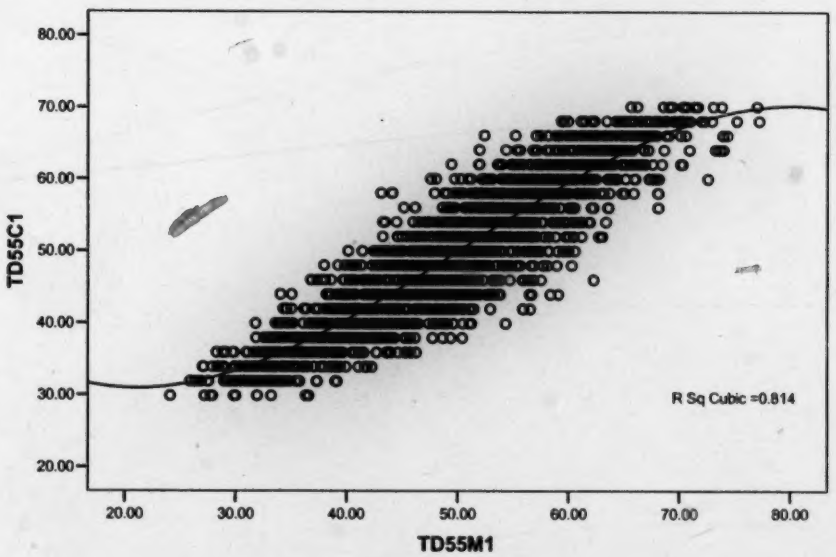


Figure 19a: Scatter plot of CTM and MIRM - Subtest 2 - 0003CS

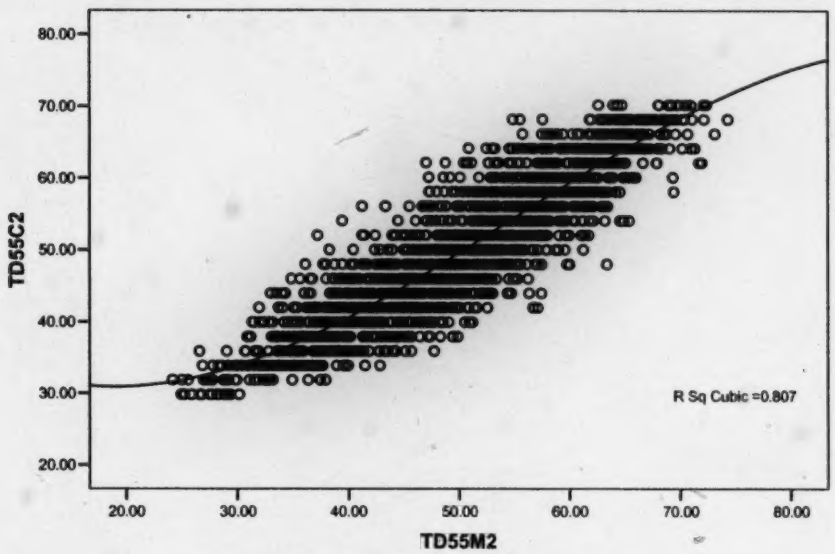


Figure 20: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0003CS

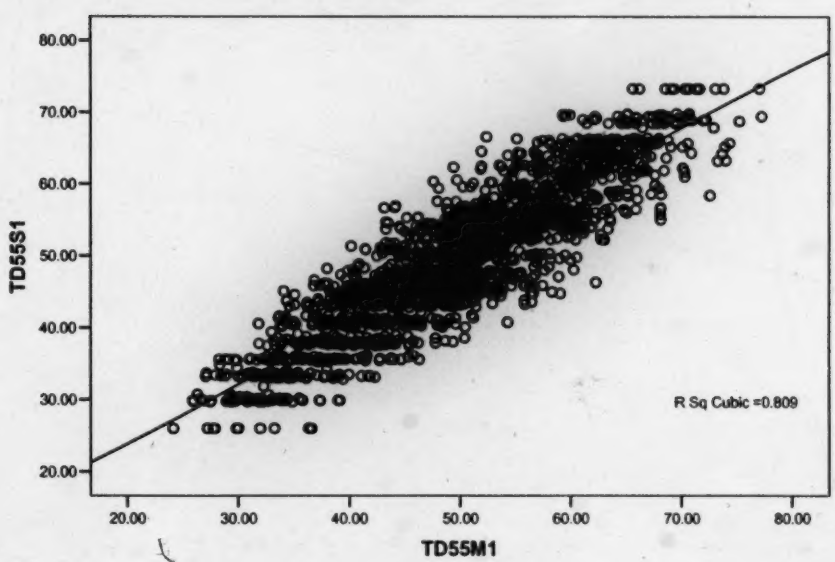


Figure 20a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0003CS

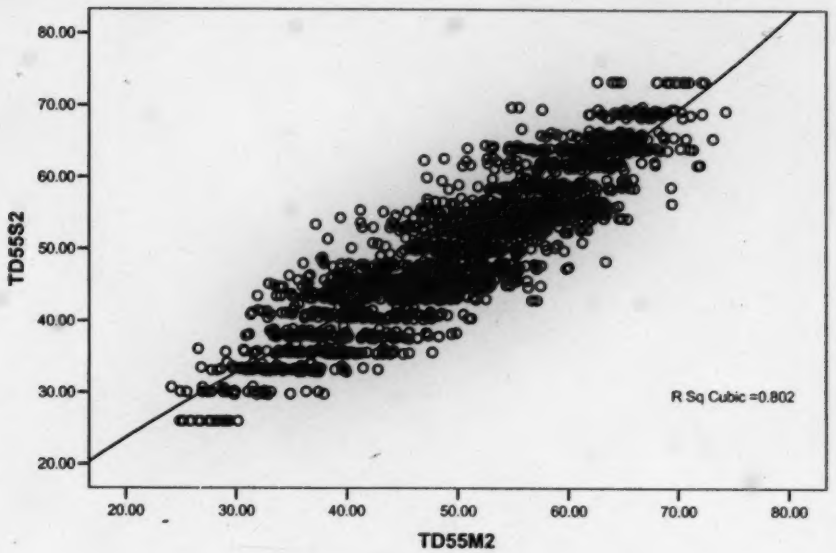


Figure 21: Scatter plot of CTM and MIRM - Subtest 1 - 0006CS

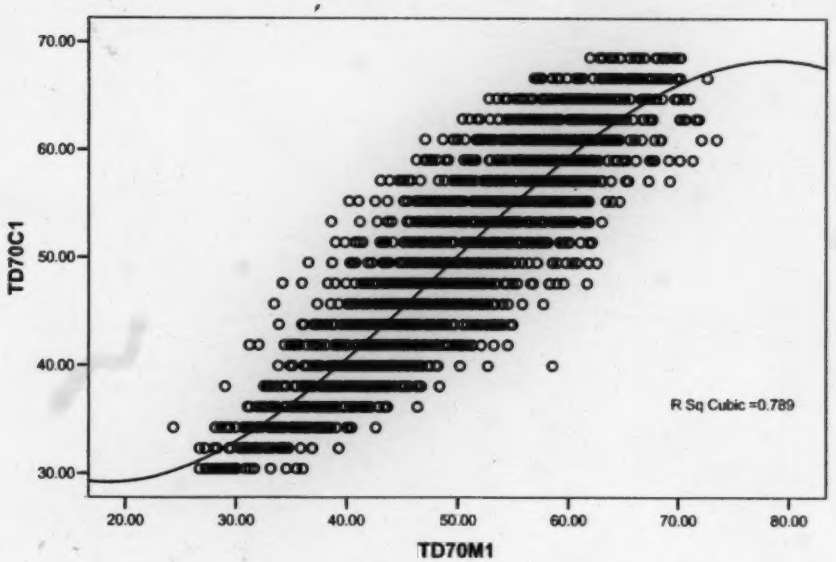


Figure 21a: Scatter plot of CTM and MIRM - Subtest 2 - 0006CS

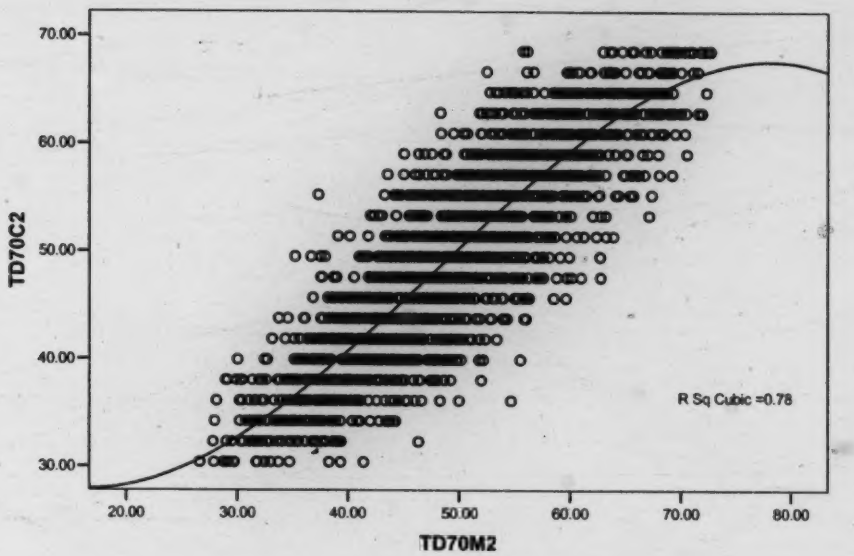


Figure 22: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0006CS

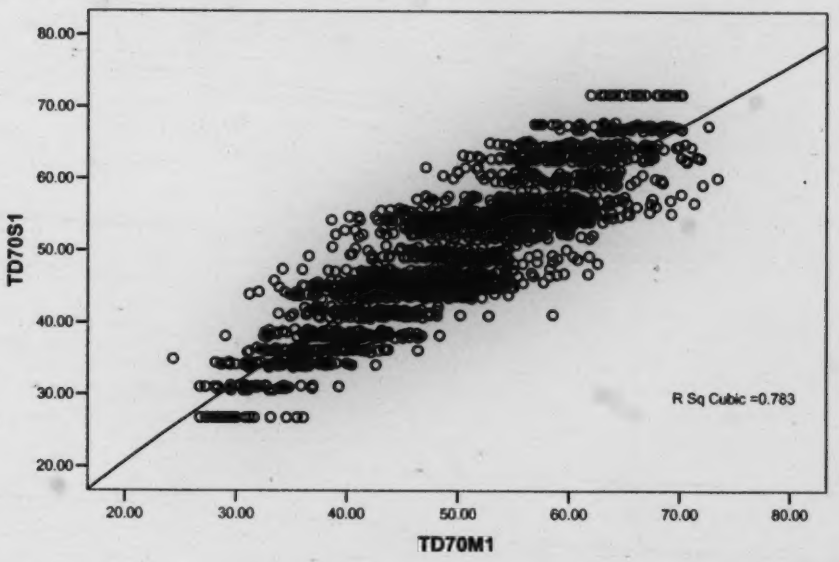


Figure 22a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0006CS

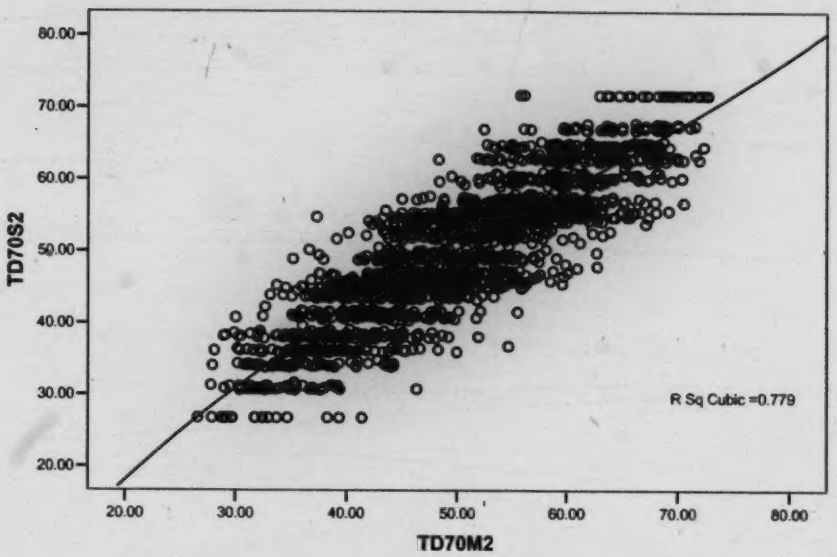


Figure 23: Scatter plot of CTM and MIRM - Subtest 1 - 0-100CS

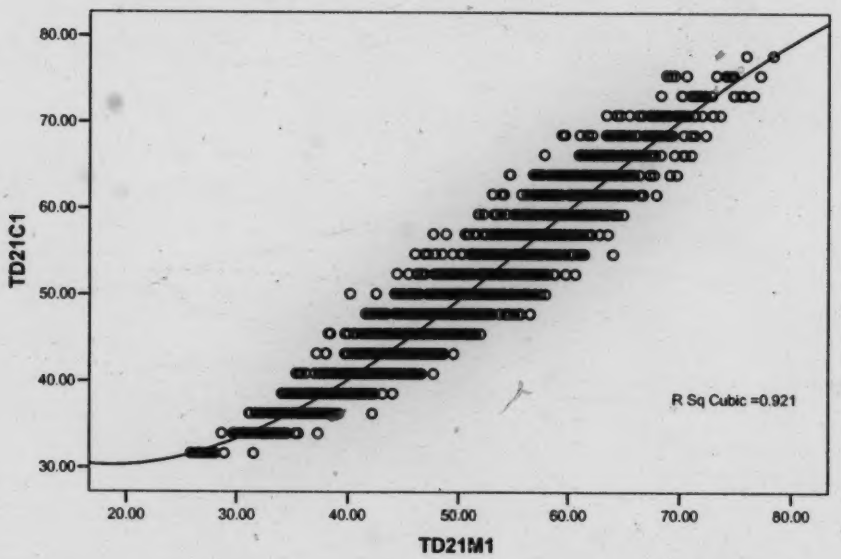


Figure 23a: Scatter plot of CTM and MIRM - Subtest 2 - 0-100CS

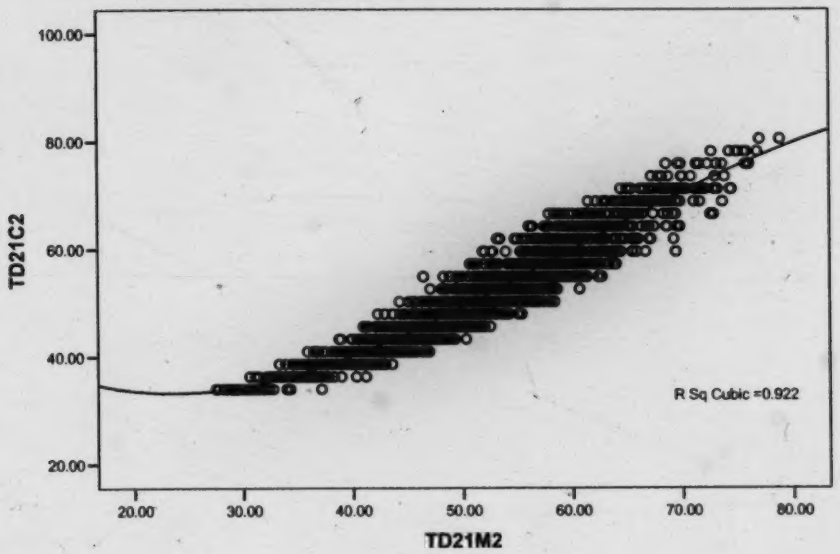


Figure 24: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-100CS

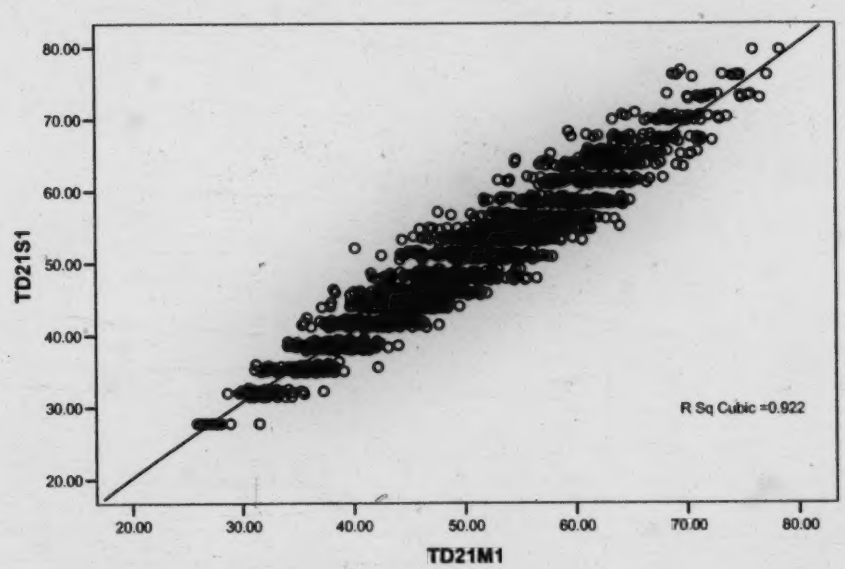


Figure 24a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-100CS

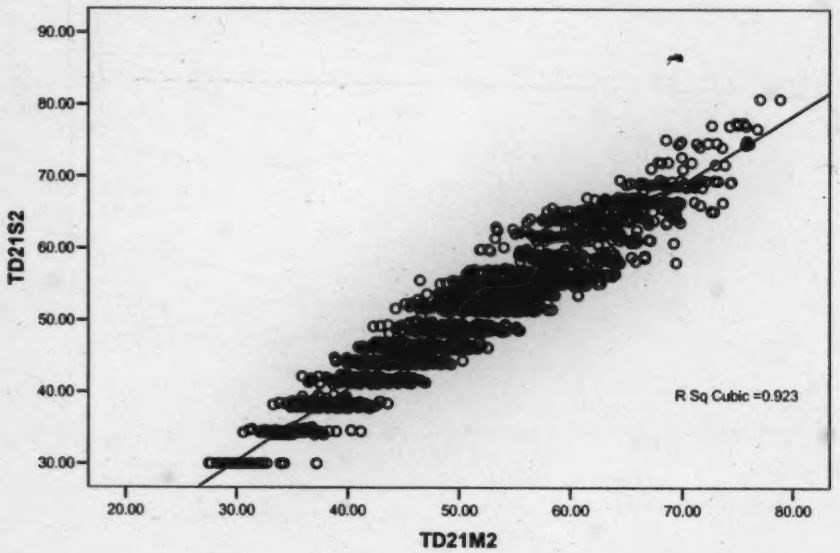


Figure 25: Scatter plot of CTM and MIRM - Subtest 1 - 0-103CS

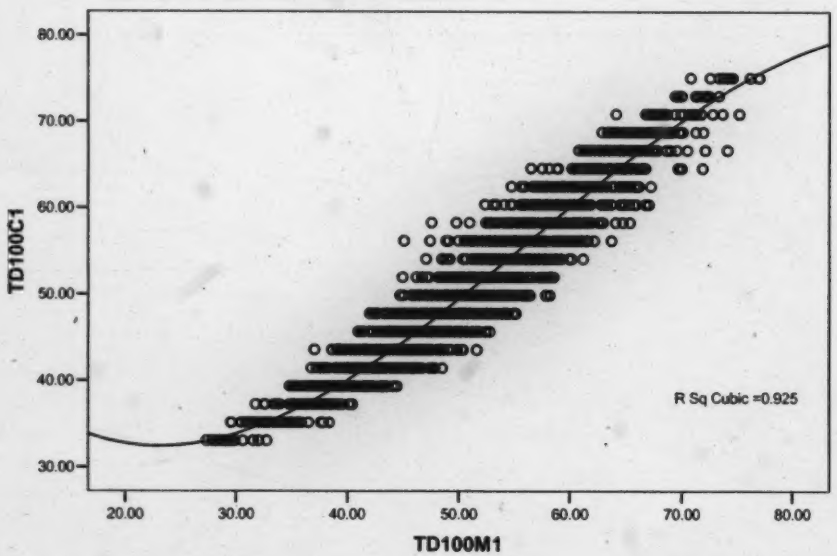


Figure 25a: Scatter plot of CTM and MIRM - Subtest 2 - 0-103CS

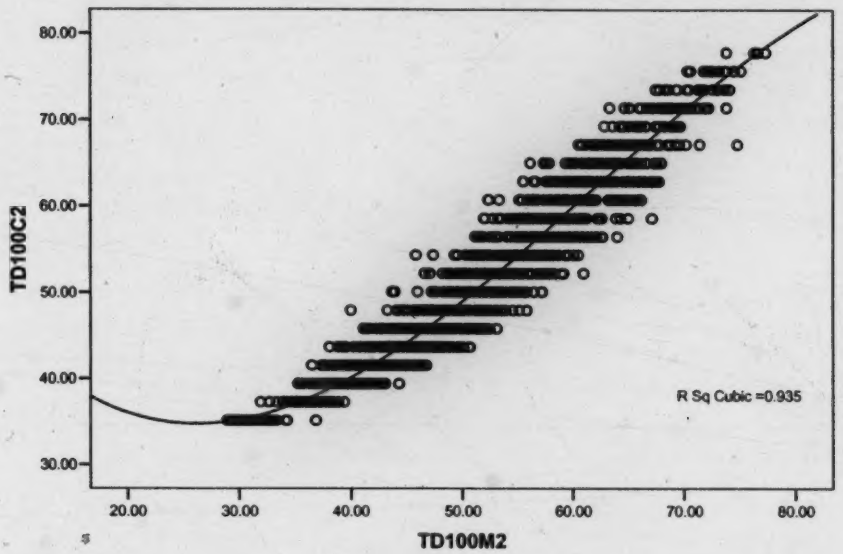


Figure 26: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-103CS

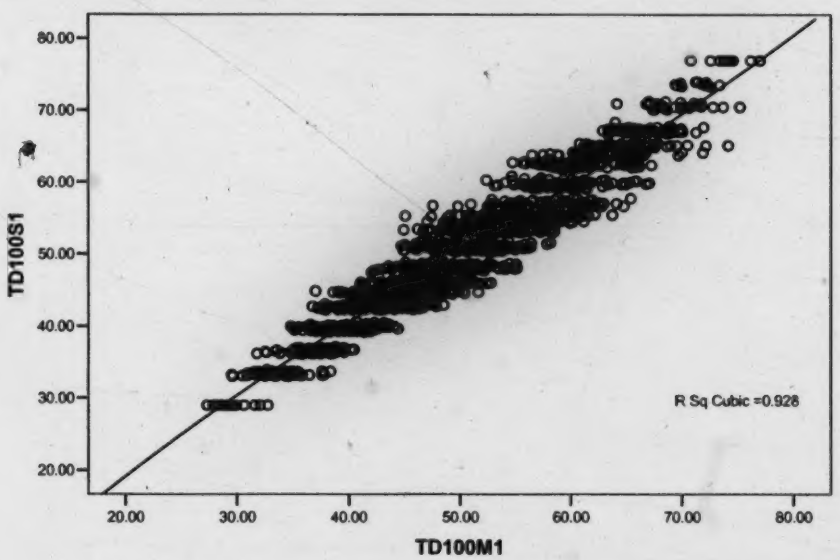


Figure 26a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-103CS

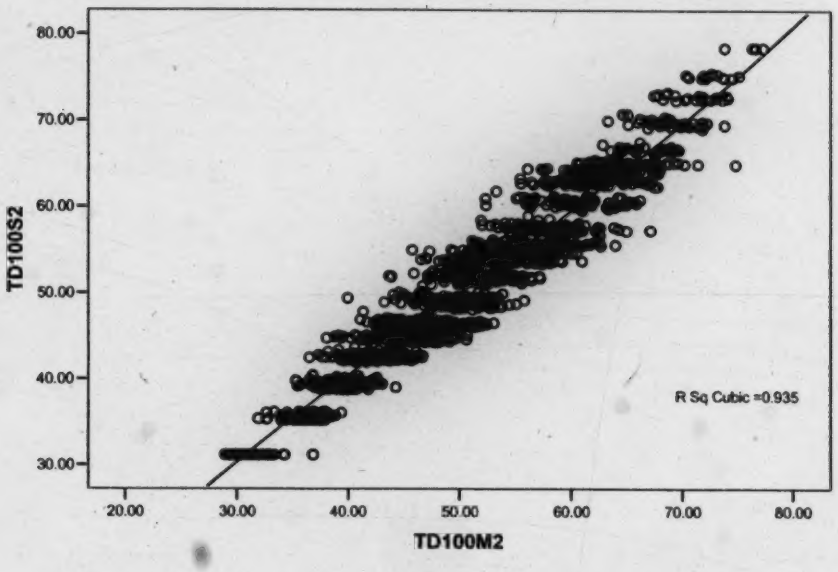


Figure 27: Scatter plot of CTM and MIRM - Subtest 1 - 0-106CS

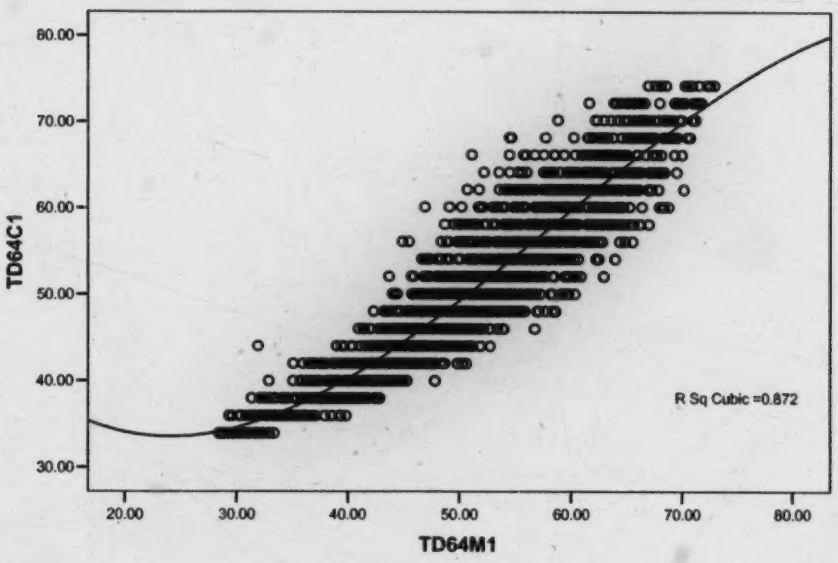


Figure 27a: Scatter plot of CTM and MIRM - Subtest 2 - 0-106CS

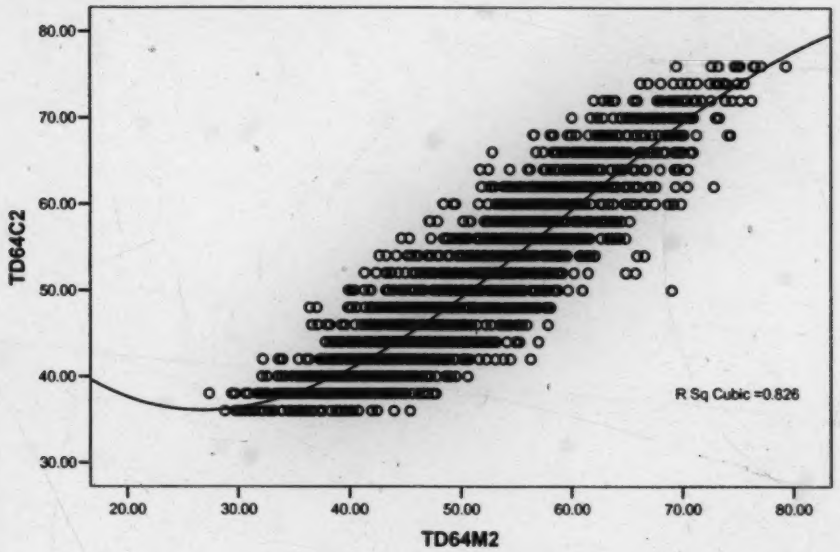


Figure 28: Scatter plot of UIRM (S) and MIRM - Subtest 1 - 0-106CS

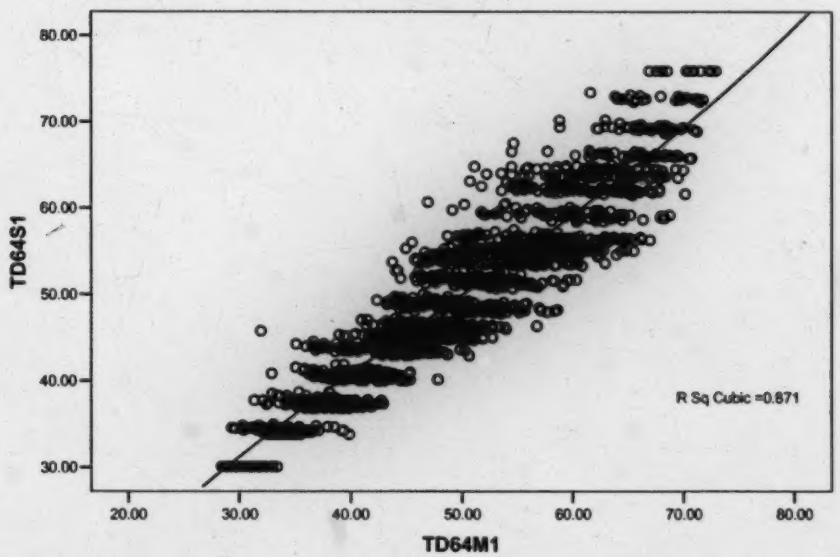
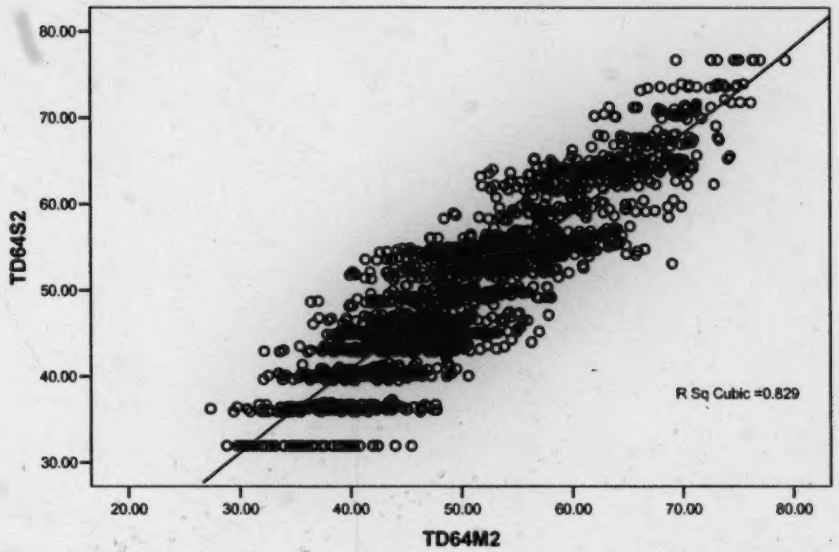


Figure 28a: Scatter plot of UIRM (S) and MIRM - Subtest 2 - 0-106CS



Appendix C

Multidimensional Item Statistics

The multidimensional discrimination parameter, the multidimensional difficulty, and the angular directions are presented in this section for both simple and complex structure. Table 1c depicts the MDISC, MDIFF, and the angles in degrees for each item when the test structure is simple. Table 2c depicts the MDISC, MDIFF, and the angles in degrees for each item when the test structure is complex.

Table 1c

Multidimensional Item Statistics for Simple Structure

				Statistics			
Item	a1	a2	d	MDISC	D	Radian	Angle
1	0.60	0.21	-1.00	0.63	1.58	0.33	19
2	0.60	0.20	-0.90	0.63	1.43	0.31	18
3	0.60	0.18	-0.80	0.63	1.28	0.30	17
4	0.60	0.17	-0.70	0.62	1.12	0.28	16
5	0.60	0.16	-0.60	0.62	0.97	0.26	15
6	0.60	0.15	-0.50	0.62	0.81	0.24	14
7	0.60	0.14	-0.40	0.62	0.65	0.23	13
8	0.60	0.13	-0.30	0.61	0.49	0.21	12
9	0.60	0.12	-0.20	0.61	0.33	0.19	11
10	0.60	0.11	-0.10	0.61	0.16	0.17	10
11	0.60	0.10	0.10	0.61	-0.16	0.16	9
12	0.60	0.08	0.20	0.61	-0.33	0.14	8
13	0.60	0.07	0.30	0.60	-0.50	0.12	7
14	0.60	0.06	0.40	0.60	-0.66	0.10	6
15	0.60	0.05	0.50	0.60	-0.83	0.09	5
16	0.60	0.04	0.60	0.60	-1.00	0.07	4
17	0.60	0.03	0.70	0.60	-1.17	0.05	3
18	0.60	0.02	0.80	0.60	-1.33	0.03	2
19	0.60	0.01	0.90	0.60	-1.50	0.02	1
20	0.60	0.00	1.00	0.60	-1.67	0.00	0
21	0.21	0.60	-1.00	0.63	1.58	1.24	71
22	0.20	0.60	-0.90	0.63	1.43	1.26	72
23	0.18	0.60	-0.80	0.63	1.28	1.27	73
24	0.17	0.60	-0.70	0.62	1.12	1.29	74
25	0.16	0.60	-0.60	0.62	0.97	1.31	75
26	0.15	0.60	-0.50	0.62	0.81	1.33	76
27	0.14	0.60	-0.40	0.62	0.65	1.34	77
28	0.13	0.60	-0.30	0.61	0.49	1.36	78
29	0.12	0.60	-0.20	0.61	0.33	1.38	79
30	0.11	0.60	-0.10	0.61	0.16	1.40	80
31	0.10	0.60	0.10	0.61	-0.16	1.41	81
32	0.08	0.60	0.20	0.61	-0.33	1.43	82
33	0.07	0.60	0.30	0.60	-0.50	1.45	83
34	0.06	0.60	0.40	0.60	-0.66	1.47	84
35	0.05	0.60	0.50	0.60	-0.83	1.48	85
36	0.04	0.60	0.60	0.60	-1.00	1.50	86
37	0.03	0.60	0.70	0.60	-1.17	1.52	87
38	0.02	0.60	0.80	0.60	-1.33	1.54	88
39	0.01	0.60	0.90	0.60	-1.50	1.55	89
40	0.00	0.60	1.00	0.60	-1.67	1.57	90

Table 2c

Multidimensional Item Statistics for Complex Structure

				Statistics			
Item	a1	a2	d	MDISC	D	Radian	Angle
1	0.60	0.58	-1.00	0.83	1.20	0.77	44
2	0.60	0.56	-0.90	0.82	1.10	0.75	43
3	0.60	0.54	-0.80	0.81	0.99	0.73	42
4	0.60	0.52	-0.70	0.79	0.88	0.72	41
5	0.60	0.50	-0.60	0.78	0.77	0.70	40
6	0.60	0.49	-0.50	0.77	0.65	0.68	39
7	0.60	0.47	-0.40	0.76	0.53	0.66	38
8	0.60	0.45	-0.30	0.75	0.40	0.65	37
9	0.60	0.44	-0.20	0.74	0.27	0.63	36
10	0.60	0.42	-0.10	0.73	0.14	0.61	35
11	0.60	0.40	0.10	0.72	-0.14	0.59	34
12	0.60	0.39	0.20	0.72	-0.28	0.58	33
13	0.60	0.37	0.30	0.71	-0.42	0.56	32
14	0.60	0.36	0.40	0.70	-0.57	0.54	31
15	0.60	0.35	0.50	0.69	-0.72	0.52	30
16	0.60	0.33	0.60	0.69	-0.87	0.51	29
17	0.60	0.32	0.70	0.68	-1.03	0.49	28
18	0.60	0.31	0.80	0.67	-1.19	0.47	27
19	0.60	0.29	0.90	0.67	-1.35	0.45	26
20	0.60	0.28	1.00	0.66	-1.51	0.44	25
21	0.58	0.60	-1.00	0.83	1.20	0.80	46
22	0.56	0.60	-0.90	0.82	1.10	0.82	47
23	0.54	0.60	-0.80	0.81	0.99	0.84	48
24	0.52	0.60	-0.70	0.79	0.88	0.86	49
25	0.50	0.60	-0.60	0.78	0.77	0.87	50
26	0.49	0.60	-0.50	0.77	0.65	0.89	51
27	0.47	0.60	-0.40	0.76	0.53	0.91	52
28	0.45	0.60	-0.30	0.75	0.40	0.92	53
29	0.44	0.60	-0.20	0.74	0.27	0.94	54
30	0.42	0.60	-0.10	0.73	0.14	0.96	55
31	0.40	0.60	0.10	0.72	-0.14	0.98	56
32	0.39	0.60	0.20	0.72	-0.28	0.99	57
33	0.37	0.60	0.30	0.71	-0.42	1.01	58
34	0.36	0.60	0.40	0.70	-0.57	1.03	59
35	0.35	0.60	0.50	0.69	-0.72	1.05	60
36	0.33	0.60	0.60	0.69	-0.87	1.06	61
37	0.32	0.60	0.70	0.68	-1.03	1.08	62
38	0.31	0.60	0.80	0.67	-1.19	1.10	63
39	0.29	0.60	0.90	0.67	-1.35	1.12	64
40	0.28	0.60	1.00	0.66	-1.51	1.13	65